The Dissertation Committee for Hyunjin Shin Certifies that this is the approved version of the following dissertation:

Algorithms for Biomarker Identification Utilizing MALDI TOF Mass Spectrometry

Committee:

Jonathan W. Valvano, Supervisor

Mia K. Markey, Co-supervisor

Joydeep Ghosh

Alan C. Bovik

Edward Marcotte

Keith A. Baggerly

Algorithms for Biomarker Identification Utilizing MALDI TOF Mass Spectrometry

by

Hyunjin Shin, B.S., M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2006

Dedicated to

God

My parents Young-Kyun Shin and Bookang Rhee

My sister Hyunjung Shin

Acknowledgements

"Some trust in chariots and some in horses, but we trust in the name of the Lord our God." – PSALM 20:7

First, I truly confess that all my doctoral work could not have been done without God's guidance. He has always been with me and given me wisdom during my doctoral study. I give all glory to God alone.

I would like to give thanks to my parents for their consistent support, encouragement, and prayer. Their love for me is the most important source that enabled me to continue my doctoral study. I would like to thank my father for his constant advice and encouragement in spite of all his personal difficulties over the past few years. I will never forget my mother's prayers for me every early morning. The fact that my mother always prays for me has encouraged me to go forward without giving up. Also, I would like to thank my dear sister for her wise advice. She has always listened to me very carefully and has given advice that I needed at the moment. I am also grateful for all the prayers and wishes from my grandparents and other relatives. Particularly, my grandmother's prayers have been great strength to me.

I would like to give thanks to my supervisor Prof. Jonathan W. Valvano for his sincere advising. Discussions with him helped me to correctly understand the key of the problems with my research. Particularly, I truly appreciate his help with my noise analysis of MALDI TOF mass spectrometry. As my supervisor, he also has taken good

care of all the work related to earning my Ph.D. degree. He has been more than only a supervisor to me. Thank you Prof. Valvano.

I would like to give my thanks to my co-supervisor Prof. Mia K. Markey. Because she is not an ECE GSC member, she could not be my official supervisor. However, she has given me far more than any other supervisors could give to their students. Everything that I have now is what I have learnt from her. Without her, I could not be called as Dr. Shin that I am now. In the acknowledgements of her dissertation, Mia said that she would be very proud if she provided her own first student with anything close to the quality and depth of mentoring that her advisor had given her. I do not know how much Mia received from her advisor, but as her first student, I do know that she should be proud now and she deserves it. I also wish to be an excellent supervisor like Mia to my students in the future.

Many thanks are due to my dissertation committee: Prof. Joydeep Ghosh, Prof. Alan C. Bovik, Prof. Edward Marcotte, and Prof. Keith A. Baggerly. Their advice has been of great help for me to complete my doctoral research. I would also like to thank my collaborator Dr. John M. Koomen for his advice and provision of mass spectrometry noise data. Thanks to him, I could quickly understand the fundamentals of MALDI TOF and could perform my noise analysis study.

I also want to thank my BMIL lab members, Min, Shalini, Mehul, Wendy, Sun, Qiu, and Ernest. I will never forget the time with them. They have been my good friends and peer mentors. The time with my graduate and undergraduate assistants should be appreciated. I want to thank Preethi, Pierre, Bryan, Maria, Sheldon, and Hyewon for their inspiration and personal help. I also appreciate the technical assistance from our great computer administrators, Al, Zack, and Chris. I thank the BMIL lackeys, Vince, Charles, Lesley, and Khuyen for their Endnote work. I would also like to thank the staff members

of BME and ECE departments, Joni, Ann, Heidi, Majid, Cheryl, Maggi, and Melanie for their administrative work.

I would like to thank my church members. I would like to give my special thanks to Pastor Kookjin Nam and his wife, Pastor Ilsun Kim and his wife, Elder Ryan Kang and his wife, Deacon Taejung Yoon, Yulyung, Eunha, Chulho, Eun-a, Ilyoung, Undai, Hyojin, and Hakbum for their constant prayers. I want to share my joy with my old friends, GS Lee, Seokyoon, Jaeowook, and Kwanghun (DBS). I would also like to thank all the PSALM members for their persistent prayers and love.

Last but not least, I want to see my old friends in Korea and say "thank you for being my friends." I remember Jaeyoon, Sangmok, Hyungjo, Joonsung, Seungbok, Sunggon, and Sanghoon. They have been good friends to me for decades. I really miss my friends, Chaeyoon, Jungmin, Eusik, Jangwon, Yoonjin, Kwanghae, and Jihun (Josammosa). I wish that everything would go well with them. Finally, I want to see my old lab mates, Icaro, Ling, Allen, and Chi-chung.

Algorithms for Biomarker Identification Utilizing MALDI TOF Mass Spectrometry

Publication No._____

Hyunjin Shin, Ph.D.

The University of Texas at Austin, 2006

Supervisor: Jonathan W. Valvano

Co-supervisor: Mia K. Markey

Currently, the best way to reduce the mortality of cancer is to detect and treat it in the earliest stages. Technological advances in genomics and proteomics have opened a new realm of methods for early detection that show potential to overcome the drawbacks of current strategies. In particular, pattern analysis of mass spectra of blood samples has attracted attention as an approach to identification of potential biomarkers for early detection of cancer. Mass spectrometry provides rapid and precise measurements of the sizes and relative abundances of the proteins present in a complex biological/chemical mixture. However, this high-throughput nature of mass spectrometry has also raised a need for the development of efficient and effective bioinformatics tools for finding biologically meaningful information. Many scholars are interested in preprocessing of raw mass spectra and in extracting and selecting features from preprocessed mass spectra. These are key issues for accurate biomarker identification. Thus, in order to improve the process of biomarker identification using mass spectrometry, I have postulated a noise

model for MALDI TOF mass spectrometry from the perspective of stochastic signal processing, and have attempted to measure the spectral characteristics of components in the noise model. Noise in mass spectrometry can interfere with identification of the biochemical substances in a sample. I assumed that the noise in MALDI TOF mass spectrometry is composed of three components: noise from instrumentation, noise from random ion motions, and chemical noise. In this dissertation, I have separated and analyzed noise from instrumentation and chemical noise using parametric power spectral density estimation and wavelet-based analysis, respectively. In addition to these noise analysis studies, I also have designed an algorithm that can select independent and discriminant features from mass spectra of complex protein samples by reducing redundant and irrelevant information.

Table of Contents

List of Tables

List of Figures

Chapter 1: INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

Cancer is a major public health concern in the U.S. In 2006, there will be about 1.4 million new cancer cases and more than 564,000 deaths due to cancer [1, 2]. Cancer accounts for one of every four deaths in the U.S. [1]. Currently, the best way of reducing the mortality of cancer is to detect and treat it in the earliest stages [3]. For example, when breast cancer is detected at the advanced stage, in which cancer is metastasized from the original organ to others, the survival rate is only 26%. However, when breast cancer is detected at the early stage, in which cancer is localized in organ of origin, the survival rate increases to 98% [1]. Similarly, the survival rate of prostate cancer soars from 34% when the cancer is detected at the advanced stage to nearly 100% at the early stage [1].

A cancer screening test is considered efficacious if it results in a decrease in cause-specific mortality. Necessary evidence in favor of a particular screening test includes earlier detection of disease than would have occurred due to presentation of symptoms and evidence that earlier treatment will result in a better outcome. (There is a helpful overview online at http://cancer.gov/cancertopics/pdq/screening/overview). Screening and diagnostic tests are typically evaluated in terms of their sensitivity and specificity. Sensitivity is the fraction of disease cases that are correctly identified as disease. Specificity is the fraction of non-disease cases that are correctly identified as non-disease.

Currently, there exist effective screening tests for use in the general population for only a few types of cancer. The screening methods that are best supported by the evidence to date are (1) the Pap smear for cervical cancer screening, (2) mammography

for breast cancer detection, and (3) fecal occult blood testing for colorectal cancer screening. While there are limitations to each of these methods, there is evidence that they have made substantial contributions to reducing the morbidity and mortality due to cancer.

A Pap smear is an exfoliative cytological staining procedure that can help identify premalignant and malignant changes in the cervical epithelium. The incidence of, and mortality of women due to, cervical cancer has declined about 70% in the U.S. since the Pap was introduced in the 1950s. Use of this screening test reduces the incidence as well as mortality since the Pap smear can detect precancerous changes that can be treated. However, with a specificity of only 63%, many false-positive Pap smears occur in screening the general population, in which cervical cancers and precancerous lesions are thankfully rare [4]. Unfortunately, false negative Pap smears also occur since the sensitivity of the exam is 73% [4].

Mammography, x-ray imaging of the breasts, is used to detect breast cancer. Mammography has reduced the mortality of breast cancer by approximately 25-30% in the U.S. since the 1970s [5, 6]. Mammography also suffers from false positives due to the combination of moderate specificity and low disease prevalence. Only 15-34% of the positive cases from mammography are found to be actually malignant at biopsy [7, 8]. False negatives also occur since the sensitivity of mammography is approximately 90% [9].

Fecal occult blood testing (FOBT) is used for the early detection of colorectal cancer. It can detect colorectal cancer by measuring blood loss in the stool, which mainly occurs due to colorectal neoplasms [10, 11]. FOBT is reported to have reduced the mortality of colorectal cancer in the U.S. by 33% [10-13]. FOBT has a fairly high specificity of 96-98% [13]. However, because the sensitivity of the FOBT is merely 40%

[13], there is concern that the diagnosis and treatment of colorectal cancer can be delayed due to false negative tests.

An ideal cancer screening method would be accurate, non-invasive, and inexpensive. As discussed above, the accuracy levels of existing screening methods are far from ideal. The false negatives resulting from screening methods in current use delay the diagnosis of cancer, which can lead to increased morbidity and mortality. The false positives generated by the early detection methods used in current practice necessitate additional diagnostic testing which increases costs, discomfort, and stress. Existing screening modalities are all invasive to some extent: a Pap smear is obtained from a pelvic exam, mammography is based on exposure to ionizing radiation and compression of the breasts, and FOBT requires a stool sample. Many variables are believed to impact compliance with existing screening programs, but physical discomfort and embarrassment are probably important factors (*e.g.*, [14]). The costs associated with current approaches to cancer screening remain problematic as well (*e.g.*, [15]).

Recent technological advances in genomics and proteomics have opened a new realm of early detection, showing potential to overcome the drawbacks of current early detection strategies. A biomarker is a biologically derived molecule in the body that indicates the progress or status of a disease. The concentration level or pattern of biomarkers related to a certain type of cancer can be used for early detection or diagnosis. Studies of the application of biomarkers for early cancer detection can be summarized into two categories: the usage of a single biomarker and the pattern analysis of multiple biomarkers.

When a single biomarker is used, the concentration level of the biomarker is taken as an indicator of the presence or absence of cancer. A threshold is set on the concentration level of a biomarker and if the concentration level is higher than the

threshold, the specimen is considered "positive" for cancer. An example of early detection based on a single biomarker is the use of prostate-specific antigen (PSA) in blood to detect prostate cancer. PSA is a protein secreted by the epithelial cells of the prostate gland. The PSA level in blood is generally low in healthy people or patients with benign prostate disease such as benign prostatic hyperplasia (BPH), but it tends to rise in many patients with malignancies [16]. However, the specificity of using the concentration level of PSA as an indicator of prostate cancer ranges from only 18-50% with a sensitivity of 70-90% [16]. The low specificity causes many false positives to occur; therefore, unnecessary biopsies are performed to corroborate the absence of prostate cancer. There is considerable debate as to whether screening for prostate cancer by PSA is efficacious [3, 17-20].

The problems encountered with the PSA biomarker suggest limitations that may plague any test based on a single biomarker. Given the high level of biological variability and the fact that cancer cells are derived from normal cells in the body, it may not be possible to identify a single circulating protein that can identify the presence of cancer with high sensitivity and specificity in the general population. Even for high-risk populations (*e.g.*, CA 125 for women at high risk for ovarian cancer [21]), it is unlikely that single biomarkers will provide as accurate testing as the use of multiple biomarkers.

An important difficulty in developing tests based on single biomarkers is that the identification process demands a vast amount of time and labor [19]. Traditionally, 2-D gel electrophoresis (2DE) has been used for biomarker identification in tandem with mass spectrometry [17, 19, 22-24]. A protein expressed differently between cancer and normal specimens is extracted using 2DE and the extracted protein is identified by peptide fingerprinting using mass spectrometry and protein/peptide databases. 2DE is the

bottleneck of this process because it is extremely time-consuming and laborious [19, 25, 26].

Recently, pattern analysis of multiple biomarkers in blood samples has attracted attention as an alternative to the usage of a single biomarker for early detection of cancer. The pattern differences of protein profiles between cancer and healthy samples are perceived using data mining algorithms. Multiple proteins rather than a single protein are used as a 'panel' of biomarkers in this approach. Because these pattern differences originate from the complexity of blood, which is a mixture of thousands of proteins, a protein profiling modality with high-throughput and high sensitivity is required.

Mass spectrometry has the potential to meet these requirements by providing the sizes and relative abundances of the proteins in a complex biological/chemical mixture in a rapid and precise manner [27-31]. Recently, studies have been performed on a several types of cancer, including ovarian [32-46], prostate [41, 47-57], breast [58-60], bladder [61, 62], lung [63-78], liver [79], pancreatic [80-85], renal cell carcinoma [86], colorectal [87], leukemia[88], and astroglial tumor [89]. Most of these studies reported fairly high sensitivities and specificities (over 80%). However, many questions have been raised about the reliability of these reported results due to the "black box" methods employed [52, 90-94].

Diamandis pointed out that the peak height does not linearly correspond to the protein abundance because mass spectrometry only provides the *relative* abundance of proteins in a sample [92]. He also inquired about why different data mining algorithms had produced different sets of potential biomarkers. He took as an example the studies on prostate cancer performed by Qu *et al*. [53] and Petricoin *et al*. [95]. They achieved high sensitivities (96%: Qu *et al*., 95%:Petricoin *et al*.) and specificities (98%: Qu *et al*., 83%:Petricoin *et al*.) with different sets of potential biomarkers selected through different

data mining algorithms. Another question is why known biomarkers, for example PSA, do not seem to be reflected by the studies so far and the potential biomarkers found in these studies have fairly low mass [90-92]. Since low mass proteins are easily cleared by the kidney, the efficacy of a panel of low mass proteins appears to be suspicious. Related to this, Diamandis and Merwe also raised another question on whether or not the putative biomarkers identified through the "black box" methods originate from cancer-specific pathological states in the body [94]. They took an example Koomen *et al.*'s study on the identification of potential biomarkers for pancreatic cancer [84]. Koomen *et al.* identified several biomarker candidates for pancreatic cancer from mass spectra of human plasma of healthy people and pancreatic cancer patients using statistical and biochemical tests [84]. However, Diamandis and Merwe argued that these biomarker candidates can be only high abundance non-cancer-specific proteins in blood, which are produced by nonspecific epiphenomena of cancer presence. Moreover, they suspected that mass spectrometry technologies such as MALDI TOF or SELDI TOF are not sensitive enough to detect low abundance clinically useful biomolecules without an aid of powerful fractionation [94].

In addition to Diamandis' questions, Baggerly *et al.* [93, 96] emphasized the problems in quality control indicated by the lack of reproducibility of the studies of Petricoin *et al.* [95] and Zhu *et al.* [35]. In these both studies, the ovarian cancer data sets posted on the website of the clinical proteomics program under the National Cancer Institute (http://home.ccr.cancer.gov/ncifdaproteomics/) were analyzed to identify diagnostic signatures for ovarian cancer. Baggerly *et al.* attempted to reproduce the experimental results obtained by Petricoin *et al.* by following the proposed bioinformatic algorithms as much as possible; however, Baggerly *et al.*'s analyses imply that the apparent successes of the study may have been due to artifacts of sample processing

rather than actual biological pattern differences [93]. In the analyses on Zhu *et al.*'s study, Baggerly *et al.* also showed that the peaks identified as potential biomarkers in one data set may not have consistently occurred in another set measured on a different date from the first set [96]. Similarly, Yasui discussed the variability of the relative abundance of the same protein across chips and samples, which also points to the need for active and systematic internal quality controls [52].

Recently, some progress has been made in addressing these important questions. For example, low mass biomarkers may be more meaningful than many had believed because other high abundance and high mass proteins such as albumin can act as carriers of low mass biomarkers. These carrier proteins enable low mass biomarkers to stay in the body longer than expected [97]. Powerful fractionation techniques amplify the concentration of these low mass biomarkers by isolating them from the carrier proteins such that mass spectrometers can sufficiently detect the pathological signatures of these low mass biomarkers [18, 97].

In addition, in response to Baggerly *et al.* [96], Liotta *et al.* pointed out that the two ovarian data sets used in Zhu *et al.*'s study were measured under different experimental settings (*e.g.,* chemistries on protein chips, pH, laser energy intensity, etc.) as well as on different days; thus, simple comparisons of two different mass spectra data sets may lead to a hasty generalization about the reproducibility of the technique [98]. Similarly, Grizzle and Meleth also maintained that it would be very unlikely for different laboratories to derive similar sets of biomarker candidates when applying different bioinformatics algorithms to samples obtained from non-identical patient populations [99].

Such issues related to reproducibility can be resolved to some extent if strongly standardized calibration and instrumentation protocols are shared among laboratories.

Recently, Semmes *et al.* reported that "between-laboratory" reproducibility of SELDI TOF MS can reach "within-laboratory" reproducibility levels if calibration and instrumentation protocols are strongly standardized among different laboratories [100]. Six different institutions succeeded in classifying prostate cancer sampled from healthy samples using a classifier trained in an institution within an acceptable variance of error rates after calibrating the SELDI TOF MS machines with a standard of pooled serum samples. This study was performed as part of an on-going effort to validate the approach of cancer detection through serum protein expression profiling using SELDI TOF MS [101].

However, many questions remain unanswered. In Semmes *et al.*'s study, while mass accuracy of the healthy samples used for the quality control agreed within an acceptable variance, their peak intensities, especially small peak intensities, showed fairly high variation despite of careful calibration. Moreover, for classification, Semmes *et al.* selected prostate cancer and healthy samples that had been used in building the classifier in their previous study and on which the classifier performed well. Thus, as Semmes *et al.* discuss in their article, their study only shows the possibility that the experimental platform can be reproducible under very rigorous unified calibration and instrumentation protocols and more work is needed on this important issue.

For reliable early detection based on pattern analysis of multiple biomarkers, more rigorous and systemic approaches are needed. Since most such pattern differences in mass spectra of samples such as plasma/serum are very subtle, noise can cause false positives or false negatives in peak detection by distorting the true shape of the mass spectrum. Thus, effective noise reduction that can maximally reveal the true chemical information of mass spectra must be developed to identify truly powerful biomarkers. In addition, selecting the smallest set of discriminant features that can inform the

pathological status of samples must also follow. In this dissertation, as a step toward effective noise reduction, I propose a noise model for MALDI TOF mass spectrometry, and analyze the stochastic natures of some of the noise components of the model. In order to handle a great number of features from MALDI TOF mass spectrometry successfully, I also develop a feature selection algorithm that can consider the interrelationships between features as well as the discrimination abilities.

1.2 OVERVIEW OF THE DISSERTATION

The dissertation is composed of a total of six chapters. Chapter 1 and 2 are extended from my review paper of a machine learning perspective on biomarker identification using mass spectrometry [102]. Relative to my review paper, these chapters have been updated with the most recent information of the research area. In Chapter 2, I briefly introduce the basic principles of MALDI TOF mass spectrometry, and the use of blood samples for early cancer detection. Then, I review the literature on the development of clinical decision support systems using mass spectrometry in an organized framework from a machine learning perspective. Study design and quality control (*e.g.*, sample preparation and mass spectrometer parameter settings) are also extremely important issues because data quality, which is mostly determined by these processes, affects the overall performance of decision support systems. However, since these issues are beyond the scope of my research, I will refer the reader to other papers that have discussed the topic of study design and quality control for experiments based on protein profiling techniques [100, 103-106].

In Chapter 3, I propose a noise model for MALDI TOF mass spectrometry and characterize the noise from instrumentation, which is one of the components of the proposed noise model. A parametric power spectral density estimation was used to

develop a stochastic signal model for noise from instrumentation. This method was designed by de Waele and Broersen to obtain a more reliable and robust model from multiple signal segments of a stationary random process [107]. To study the effects from the instrumentation environment and model type, mass spectra were measured from two MALDI TOF instruments of an identical model, but located at two different places, and from a more advanced model. The power spectral density (PSD) plots of the models of noise from instrumentation from the three machines were compared. In order to evaluate the effect of noise from instrumentation on a mass spectrum, I performed a simulation of MALDI TOF mass spectrometry, assuming that 57 human plasma proteins are analyzed by MALDI TOF.

Chapter 4 describes my analysis of chemical noise, which is caused by chemical impurities like matrix material in the sample in MALDI TOF mass spectrometry. I characterized chemical noise using multiple realizations and developed algorithms based on the wavelet transform that can reduce the monotonically decreasing baseline and high frequency noise that result from chemical noise [108].

In Chapter 5, I introduce a feature selection algorithm that was designed to select more independent and powerful features among a great number of features produced by protein profiling such as MALDI TOF mass spectrometry. This algorithm narrows down the search space for meaningful features by clustering features based on their linear relationship (*i.e.,* correlation), and selecting a powerful representative from each feature cluster. I tested the efficacy of this algorithm using controlled simulation data and real protein profiles of breast cancer [109, 110], which were used in Pusztai's breast cancer chemotherapy study [111].

In Chapter 6, I summarize my findings and suggest several study areas for more effective biomarker identification using MALDI TOF or other types of mass spectrometry from a biomedical informatics perspective.

Chapter 2: BACKGROUND

2.1 MASS SPECTROMETRY

Mass spectrometry provides rapid and precise measurements of the sizes and relative abundances of the proteins present in a complex biological/chemical mixture. Here I provide a very brief overview of the technique as it is typically used for identifying cancer biomarkers from blood samples. I refer the reader to other articles for a thorough review of mass spectrometry methods [112-117].

The capabilities of a mass spectrometer are determined by its ion source, mass analyzer, and detector. Protein profiling of plasma and serum has been performed primarily with a matrix-assisted laser desorption ionization (MALDI) ion source or its derivative, the surface-enhanced laser desorption ionization (SELDI) ion source coupled to a time-of-flight (TOF) mass analyzer with a chevron microchannel plate detector. The only difference between SELDI and MALDI is the use of derivatized surfaces to capture peptides and proteins based on particular physical or biochemical characteristics prior to MALDI sample preparation and mass analysis. A brief description of MALDI TOF mass analysis is given in the following paragraphs.

To prepare proteins or peptides for MALDI mass analysis, aqueous solutions of the proteins or peptides are mixed with solutions of matrix molecules, like sinapinic acid and α-cyano-4-hydroxycinnamic acid, which are present in large molar excess compared to the proteins and peptides (10,000:1). Aliquots of this mixture are deposited on the MALDI plate and allowed to dry (this procedure is referred to as the dried droplet technique). The peptides and proteins selectively cocrystallize with the MALDI matrix as the solvent evaporates. After drying, the sample plate is introduced into the vacuum chamber of the mass spectrometer and placed in the MALDI ion source. To produce ions,

an ultraviolet laser (337 nm or 355 nm) is used to irradiate the matrix crystals. The energy from these photons is transferred into translational and vibrational energy causing desorption of matrix material containing the peptide and protein analytes. The softer process ionization of MALDI (when compared to laser desorption ionization) prevents fragmentation of the protein and peptide analytes [114]. However, ionized clusters of matrix molecules produce chemical noise, which interferes with the ion signals of interest that correspond to the peptides and proteins [118, 119].

After a delay of a few hundred nanoseconds (Wiley-McLaren time lag focusing), all ions are extracted from the source and accelerated into the TOF mass analyzer. The voltage settings in the ion source determine the range of optimized ion signal, *i.e.,* the TOF has mass-dependent focusing. The ions drift in a field free region, where they are separated based on their mass-to-charge ratios. The principle behind this separation is that the potential energy of each ion in an electric field ($U = zV$) is converted into the kinetic energy of the ion in the TOF ($E = \frac{1}{2}mv^2$). By setting these equations equal to one another, the TOF equation can be derived and rearranged to calculate the m/z value for an ion:

$$zV = \frac{1}{2}mv^2 \tag{2.1}$$

$$v = \frac{l}{t} \tag{2.2}$$

$$\frac{(2Vt^2)}{l^2} = \frac{m}{z} \tag{2.3}$$

In Eq. 2.1, z denotes the ion's amount of charge, V the electric potential that accelerates the ion, m the ion's mass, and v the ion's velocity. In Eq. 2.2, l is the length of the flight tube of the TOF mass spectrometer and t the flight time of the ion. Eq. 2.3 shows that the mass to charge ratio can be represented as a quadratic function of the flight time. Ions of the same m/z have the same flight time and thus impact the detector at the

same time. When the ion strikes the detector, a cascade of secondary electrons is released. This current is captured by an anode and converted to a voltage using a preamplifier. The resulting voltage is recorded by a digital storage oscilloscope or by a digitizer card in a computer, and the amplitude of the signal corresponds to the number of ions that struck the detector in each bin of ion flight time. Other sources of noise from physical and electrical components of the mass spectrometer are also recorded (*e.g.,* high frequency noise).

Data are recorded as plots of intensity versus flight time and displayed as intensity versus m/z, which is referred to as a mass spectrum (Figure 1). In each mass spectrum, the individual ion signals correspond to nonvolatile analytes in the original sample. In protein profiling, these ion signals primarily correspond to peptides and proteins because of the analyte specificity of the matrices described above. The mass-to-charge ratios (m/z), displayed as the x-axis, can be used to calculate the molecular weights of protein or peptide in the profile. For the analysis of complex mixtures, like plasma or serum protein fractions, MALDI TOF MS has detection sensitivity in the 0.1 to 10 picomole range and mass measurement accuracy ranging from 0.01 to 0.5%. Ion signals in different mass spectra with centroids m/z values within the mass measurement error tolerance should be considered to be the same peak (protein). Because of the complexity of the samples, which produces suppression effects, and the lack of internal and external standards for quantification, the intensity of the ion signals in a protein profile does not directly correlated to protein concentration [97]. Nonetheless, relative abundances of a particular ion signal can be determined by comparing mass spectra acquired from different samples. Thus, noise reduction and normalization schemes are critical to enable accurate statistical analysis of mass spectra.

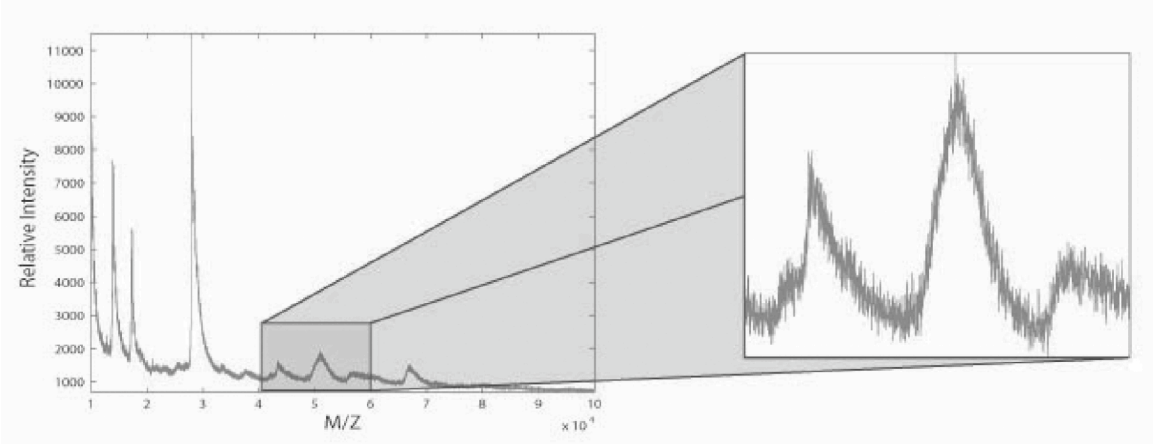

Figure 2.1: Example of a mass spectrum in which the relative abundance is plotted as a function of the mass-to-charge ratio (m/z). Notice the monotonically decreasing baseline. A portion of the spectrum has been enlarged so that the high frequency noise is apparent.

Ciphergen® (Freemont, CA) developed a SELDI TOF system to accomplish both fractionation and mass analysis in a succinct and accurate manner. SELDI TOF is a special case of MALDI TOF in which chromatography is performed using protein chips that can capture only those proteins that biochemically/chemically match certain binding characteristics (*e.g.*, hydrophobic), even when a variety of proteins are mixed together in high concentrations [29, 120-124]. The selected "fraction" of proteins deposited on the protein chip is analyzed through MALDI TOF mass spectrometry. SELDI TOF enables the amplification of the mass abundance information of more proteins than other types of mass spectrometry by using a protein chip with a predefined chromatographic surface [97].

Recently, more advanced types of mass spectrometry have been tested to improve the sensitivity to diagnostic patterns in protein profiling [34, 43, 55, 56]. Whereas traditional mass spectrometers provide 15,000-40,000 m/z data records, high-resolution

mass spectrometers can extend these to 350,000-400,000 m/z [34]. A hybrid quadrupole time-of-flight (QqTOF) such as the QSTAR pulsar I (Applied Biosystems, Inc., Framingham, MA, USA) is a frequently used model for this purpose. To the best of my knowledge, there have been no studies in which different types of high-resolution mass spectrometers are extensively compared and discussed. One expects that the development of more efficient and effective preprocessing and feature extraction/selection algorithms will be even more important issues for high-resolution MS than for traditional MALDI TOF or SELDI TOF because of the increase in size of each of data record.

2.2 BLOOD SAMPLES

This section reviews approaches that are being explored for cancer diagnosis using mass spectrometry of blood samples. There are several advantages to using blood samples because blood is readily accessible. While more invasive than some diagnostic imaging modalities, there is relatively little discomfort and low risk of side effects or adverse events associated with blood testing. Obtaining blood samples is less expensive than many other procedures. The primary disadvantage of using blood samples is that one expects that tumors located in most organs of the body will produce few proteins that will circulate in the blood at an appreciable level.

Throughout my discussion, I refer generically to "blood samples"; however, the reader should note that mass spectrometry is not performed on whole blood but on derived products, particularly plasma or serum. Plasma is the liquid portion of blood in which the cells are suspended; serum is the fluid that remains after clotting proteins are removed from plasma [125]. The advantage of using plasma rather than serum is that it contains more proteins and that the protease activity, which leads to protein degradation,

is inhibited in plasma but not in serum. However, the disadvantage of using plasma is that low abundance proteins associated with disease may be difficult to detect in the presence of a large amounts of common proteins involved in clotting. Both plasma and serum have been used in studies of cancer diagnosis using mass spectrometry and it is not yet known which is best for this kind of analysis.

There have been many studies of the serum/plasma proteomes using techniques such as 2D gel electrophoresis (*e.g.*, [126]). However, to the best of my knowledge, this information has not been incorporated into studies of cancer diagnosis using mass spectrometry. It is possible that more accurate models for sample classification could be developed if prior knowledge of blood proteins could be properly taken into account.

2.3 FRAMEWORK FOR SYSTEM DEVELOPMENT

I employ a machine learning framework to review the literature on the development of clinical decision support systems utilizing mass spectrometry of blood samples. There are five stages of data analysis in this framework. First, the spectra are preprocessed to reduce the contribution of noise and to normalize the spectra from different samples such that they are comparable. Second, features reflecting the pathological status of a sample are extracted from the mass spectra. Interpretable features, such as peaks corresponding to distinct protein species, are generally preferred. Third, highly discriminant features are selected to reduce the dimensionality of the data, which increases the likelihood of successful classification. Fourth, machine learning models are designed to distinguish cancer from normal samples based on the selected features. Fifth, the system is evaluated in terms of clinically relevant metrics such as sensitivity and specificity. Ideally, separate data sets should be used for each stage. However, in practice some form of data partitioning of a single data set, such as cross-validation or bootstrap

sampling, is employed due to the difficulties of obtaining a large number of spectra. The five stages are mutually dependent and the best combination of methods to be used at each stage must be determined empirically.

2.3.1 Preprocessing

Biomedical data are notoriously complex and variable. The goal of preprocessing methods is to "clean up" the data such that machine learning algorithms will be able to tease out key information and correctly classify new samples based on a limited set of examples. In analyzing mass spectra of blood samples, the preprocessing stage includes two main tasks: noise reduction and normalization.

In mass spectrometry, the noise is the undesired interfering signal caused by sources unrelated to the biochemical nature of the sample being analyzed and the signal is the relative abundance of ions originating from the proteins in the sample. Many studies to date have not employed explicit noise reduction schemes other than basic noise reduction methods implemented on commercial mass spectrometers (*e.g.,* the SELDI-TOF mass spectrometer from Ciphergen®, Freemont, CA). However, some investigators have explored methods for reducing noise, particularly the baseline and high frequency noise [35, 63, 67, 68, 71, 73, 103, 127-129].

Mass spectra exhibit a monotonically decreasing baseline (Figure 1). As described above, it is necessary to add a matrix material to the sample of interest. However, it is possible for the matrix material to interact with itself as well as with the sample proteins. The baseline originates from small clusters of matrix material. Because the chances of cluster formation decrease with cluster size, the baseline diminishes monotonically as the mass-to-charge ratio increases [117, 119]. The monotonically decreasing baseline can be regarded as low frequency noise because the baseline lies over a fairly long mass-to-

charge ratio range [130]. Most studies that have employed a baseline reduction method have taken a two-step approach: baseline estimation followed by subtraction of the estimated baseline from the original mass spectrum.

A variety of approaches have been explored to estimate the baseline from mass spectra. Such approaches can be summarized into two major categories: heuristic or model-based. Heuristic approaches form non-parametric estimates of the baseline from a set of mass spectra. Model-based approaches build a mathematical model of the baseline based on the physics of the mass spectrometer and estimate the parameters of the model from a set of mass spectra. The baseline estimated by either approach is then subtracted from the original spectrum. So far, there have been many more studies using heuristic approaches [63, 67, 68, 71, 84, 103] than model-based approaches [129].

There have been several studies in which a heuristic approach was used to estimate and eliminate the baseline. A local average or minimum intensity within a moving window has been used as a local estimator of the baseline and the overall baseline is estimated by sliding the window over the mass spectrum [71]. Piecewise linear regression has been applied to the regions with a monotonically decreasing baseline [67, 68]. The baseline has also been estimated by calculating the convex hull of the intensities of the proteins in a region [63]. All these algorithms seem to effectively estimate the underlying baseline, at least in some circumstances. However, the parameters of these algorithms, *e.g.*, the width of the window in a piecewise linear regression model, have been determined in an *ad hoc* manner. For methods in which a sliding window or piecewise linear regression are employed for baseline elimination, the window size is a critical factor determining the overall performance. If the window size is too large, these methods may oversimplify the curvature of the baseline with a long

straight line. If the window size is too small, they may produce an overly complex estimate of the baseline, which is very sensitive to high frequency noise.

There are no absolute standards for deciding which one among the heuristic baseline estimation algorithms is more effective than the others; each algorithm has its strengths and weaknesses. For example, choosing the minimum peak intensity within the sliding window as a local baseline estimator is superior to piecewise linear regression in terms of computation time. However, the latter method is expected to be relatively less sensitive to high frequency noise than the former one because a straight line with the minimum sum of errors between the line and peak intensities within the windows is calculated as a local estimator by linear regression. The convex hull is defined as the minimal convex set of given objects [131]. Thus, the convex hull of a mass spectrum is the piecewise straight lines connecting the local minima on the spectrum. This can be easily visualized by imagining a rubber band tightly stretched to encompassing the lower side of the mass spectrum. Since the convex hull is calculated based on the local minima, it may also suffer from the interference from high frequency noise.

To the best of my knowledge, there has only been one model-based approach reported in the literature to date [129]. Malyarenko *et al*. used a model for the baseline in SELDI TOF was developed using the phenomenon of charge accumulation that decays exponentially on the ion detector [129]. Greater emphasis should be placed on model-based approaches in the future because they may be more effective with limited data sets since *a priori* knowledge is taken into account.

Mass spectra of blood samples also exhibit an additive high frequency noise component (Figure 2.1). The presence of this noise hampers both data mining algorithms and human observers in finding meaningful patterns in mass spectra. While several prior studies have explored methods for reducing the influence of this high frequency noise

[35, 63, 73, 127-129, 132], few have attempted to identify or describe the sources of this noise or to determine proper models for its statistical characteristics [103, 118, 119, 130, 133]. Moreover, to date, no study has used such noise characterization work to develop a "model-based" high frequency noise reduction scheme.

The heuristic high frequency noise reduction approaches employed most commonly in studies to date are smoothing filters [35, 63, 132], the wavelet transform (WT) [54, 77, 127], or the deconvolution filter [129]. Typical smoothing filters are the Gaussian filter [35, 132] and moving average filter [63]. These smoothing filters smear out the high frequency noise signal in the spectra by averaging the intensities within a moving window. In the case of a Gaussian filter, the intensities are weighted by a Gaussian kernel before calculating the average. Over the past decade, the WT has been frequently used for chemical/biological signal processing [134, 135]. The WT is a type of signal decomposition algorithm that allows us to view a signal as a superposition of weighted basis functions with different frequencies and time shifts. The frequency range and time location of the high frequency noise are localized using the WT. Then the high frequency noise can be effectively reduced by manipulating the weight coefficients of the basis functions [54, 73, 127, 134, 135]. The deconvolution filter reduces noise by minimizing the sum of squared errors between the desired output and filtered signal and the power of filtered noise. In this case, it is assumed that the observed signal can be modeled as the sum of the true signal and additive stationary noise [136]. Malyarenko *et al.* applied the deconvolution filter to SELDI TOF mass spectra and reported that it reduced noise and improved the resolution [129].

All of the methods have made considerable contributions to high frequency noise reduction in mass spectra. However, since no study has extensively compared the methods introduced above on the same data set, it is difficult to conclude if one method is

better than the others. Moreover, the overall performance of those high frequency noise reduction methods is highly dependent on the choice of the filter parameters (*e.g.,* the size of the sliding window or the kernel weights) and the true effectiveness of those methods is difficult to measure due to the lack of knowledge on the statistical characteristics of the signal and noise in mass spectra.

Most noise reduction approaches to date have emphasized designing filters based on empirical insight rather than rigorous statistical analysis. However, a few studies have tried to identify the noise sources in mass spectrometry and to measure the statistical characteristics of the noise [103, 118, 119, 130, 133]. Such studies are critical because the lack of information on the statistical characteristics of the true signal and the noise may lead to the design of filters that remove the desired signal or fail to remove the noise. In other words, aggressive filtering may smear out diagnostically informative patterns and insufficient filtering may leave high levels of noise in the signal. Because low abundance proteins are expected to contain diagnostically useful information, *ad hoc* noise reduction approaches may actually make it more difficult to detect differences in the spectral patterns between cancer and healthy samples. Statistical characterization of the individual noise components in mass spectrometry can provide the basis for model-based approaches to noise reduction.

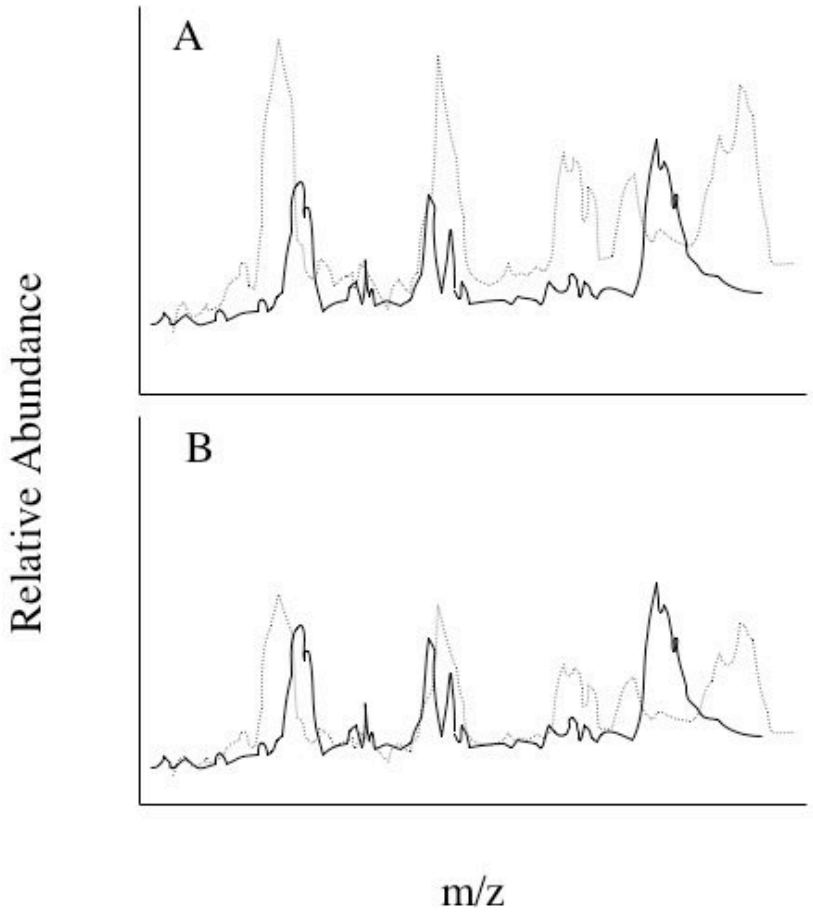

Figure 2.2: Normalization is needed in order to compare across spectra since mass spectrometry provides a measure of the relative abundance of the different proteins in a sample. In the illustration here, the original spectra (A) are normalized such that the maximum peak heights in each spectra are the same (B).

A peak in mass spectra indicates the *relative* abundance of a protein; therefore, the magnitudes of mass spectra cannot be directly compared with each other. Normalization methods scale the intensities of mass spectra in order to make mass spectra comparable (Figure 2.2). The most frequently used normalization method is normalization with respect to the total ion current (TIC), *i.e.*, the sum of all the peaks in a mass spectrum [37, 50, 58-60, 62, 71, 75, 80, 83, 137]. Normalization with respect to the mean spectrum has also been used, which is equivalent to normalization with respect to TIC [35]. Other studies have performed normalization with respect to the largest peak [66, 68] or linear scaling using the largest and smallest peak intensities [36, 38, 43, 72]. Normalization with respect to one or two peaks within a spectrum may be more sensitive to noise than normalization with respect to TIC because the effect of noise at those peaks is transferred to all other peaks through normalization while noise will be canceled out by the summation of peak intensities in normalization with respect to TIC.

The four normalization methods described above are performed within a spectrum. Normalization across samples has also been investigated. All the peak intensities at the same mass-to-charge ratio across samples can be normalized with respect to the median peak intensity [74, 79] or linearly scaled using the largest and smallest peak intensities [34, 137]. Some investigators have extended simple linear scaling by taking the peak variability into consideration [128]. These methods ignore the absolute difference in peak intensities at different mass-to-charge ratios and consider only the difference in the expression levels between cancer and normal samples. Therefore, small peaks can be considered to be as important as large peaks in normalization across samples. However, it should be noted that noise embedded in small peaks can also be amplified by such normalization methods and it still remains unanswered whether peaks belonging to different spectra can be manipulated without any precedent normalization

within a spectrum. At present, it is not clear if one normalization method is superior to the others since there have not been any studies in which normalization methods were compared on the same data set.

Some studies have investigated the use of the log transform to reduce the variability of mass spectra [39, 50, 58, 60, 69, 75, 83, 128]. However, one should be cautious in using the log transform since it may make it difficult to separate the additive noise component from the original signal. Suppose that mass spectra have additive random noise with zero mean. Such noise can easily be reduced by simple averaging; however, such noise cannot be reduced by simple averaging after a log transform because summation in the log space corresponds to multiplication in the original space. In addition to the log transform, the square root transform has also been investigated as a means of reducing the variability [75].

2.3.2 Feature extraction

Features are variables constructed from preprocessed data to summarize the properties of the data [138, 139] and the process of constructing features is called as "feature extraction". In decision support systems utilizing mass spectra, feature extraction can be defined as a process of extracting summary information reflecting the pathological status of a sample from preprocessed mass spectra.

The simplest approach to feature extraction from mass spectra is to use the abundance (intensity) information of *every* m/z measured as the features [32, 35, 36, 38-40, 64]. While this approach to feature extraction is straightforward, it places additional demand on the feature selection and classification stages since a very large number of features are used (≈15,000) and most studies employ a modest number of cases (< 500). Moreover, mass spectrometers can only distinguish the masses of proteins

within a finite resolution level. More than one m/z measured can correspond to the same protein. Thus, high levels of correlation are expected between close m/z values.

Some studies have employed binning to extract features from the raw mass spectra [43, 55, 56, 71, 72]. The m/z points are grouped into a number of bins and a feature is derived from each bin by calculating the average [72] or the maximum peak intensity [71]. The spacing of bins is usually uneven because the number of peaks is not uniformly distributed [72]. Binning is the simplest form of peak detection and alignment, which will be discussed in depth from the next paragraph, in the sense that bins are defined over the m/z axis but they are initially placed at fixed positions across multiple spectra and never adjusted again. Binning is also fairly straightforward to use; however, care must be taken in determining the size and location of bins because improper binning may lead to extraction of features, which do not reflect the pathological status of samples.

Since abundance data from within the mass error rate are considered to represent the same protein, features are often extracted from mass spectra based on the properties of "peaks" that are comprised of multiple m/z points. In this approach, feature extraction consists of three main components: peak detection, peak alignment, and calculation of feature metrics. Often, commercial software provided with mass spectrometers (*e.g.*, Ciphergen®'s SELDI TOF system) and in-house algorithms are combined in the feature extraction process.

The identification of peaks in a mass spectrum is complicated by the error in measuring the abundance as well as the mass error rate. The goal of peak detection is to identify sets of m/z values which comprise "peaks" that are higher than the noise level of a mass spectrum. In many studies, commercial software has been used to find as many peaks as possible and a predefined threshold has been applied to select peaks far higher than the noise level. For example, Ciphergen ProteinChip® software detects peaks based

on the signal to noise ratio (S/N). The S/N is an indicator of how much a peak is distinguished from background noise. If the S/N of a peak is 10, the peak has an intensity value 10 times larger than background noise. Ciphergen ProteinChip® software first selects peaks with a high signal to noise ratio (*e.g.,* S/N ≥ 10) within individual mass spectra. Then, across mass spectra, it finds more peaks with a moderately high S/N (*e.g.,* S/N ≥ 2) [127, 140]. Some researchers have explored alternative peak detection algorithms for more rigorous peak finding [52, 60, 63, 69, 103]. Most peak detection algorithms find local maxima within a certain mass-to-charge ratio range and choose the local maxima higher than a threshold of the noise level as peaks [52, 60, 68, 103]. Local maxima of a mass spectrum are located by finding the mass-to-charge ratios with the highest intensity among their N neighbors [52, 103].

Clearly, peak detection algorithms must include a definition of the noise level around a local maximum. The noise level is often defined as the average of the intensities at the mass-to-charge ratios within a moving window with a fixed size (*e.g.,* 5% of all mass-to-charge ratios in a mass spectrum) [52] or as the median elevated level from the median difference of all local maxima and their adjacent local minima in a mass spectrum [103].

Peak detection, as described above, is concerned with identifying peaks *within a single mass spectrum*. However, in order to make inferences about trends across several spectra, one must relate the peaks identified in one spectrum to the peaks identified in another spectrum. This process of matching peaks that represent the same protein specie *across multiple spectra* is referred to as "peak alignment" (Figure 2.3). In peak alignment, the peaks of multiple mass spectra within the mass error rate are grouped together and regarded as a "peak group".

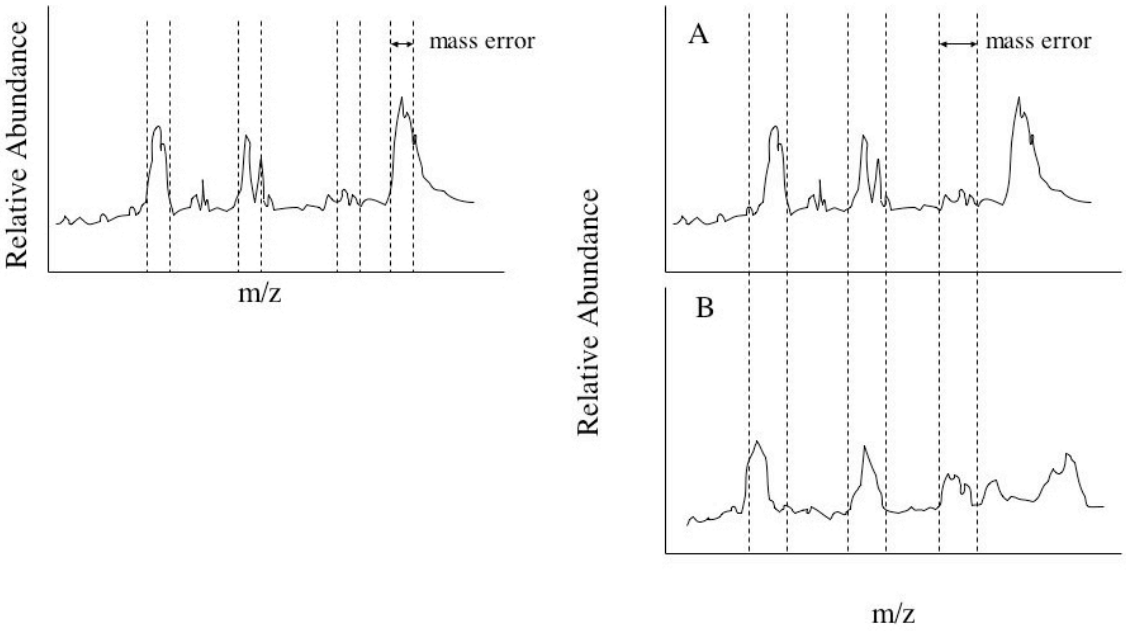

Figure 2.3: The left panel illustrates peak detection, which is concerned with identifying peaks *within a single mass spectrum*. The right panel illustrates the process of matching peaks that represent the same protein specie *across multiple spectra*, which is referred to as "peak alignment".

Most peak alignment algorithms group the peaks around a prominent peak within a moving window the size of the mass error rate in a mass spectrum. Then, the peak groups within the mass error rate are re-grouped across [37, 49-51, 60, 68, 74, 80, 83, 132]. In one study, a genetic algorithm was employed to optimize the process of window-based peak alignment [70]. Peak alignment simply based on the mass error rate can produce peak groups that cannot effectively represent proteins in a complex sample. A genetic algorithm was used to identify the peaks that were present across the most samples while at the same time avoiding those that were within the mass error rate of the peaks that were already selected.

After peak detection and peak alignment, one must define the metrics of a peak group that will serve as features. Feature metrics related to peak heights have been used in most studies. The maximum peak height [36, 67], average peak height [72, 80], and median peak height of a peak group [76] have been used. Instead of retaining the peak height as continuous feature data, binary [52] and discretized feature [66] values have also been investigated as a way to alleviate the variability of feature values across samples that can deteriorate the generalization of the classifier. Binary feature values indicate whether a peak is expressed over the noise level and further discretized feature values specify the degree to which a peak is expressed. Some studies have employed the sum of peaks in a peak group, *i.e.*, the ion current of a peak group, in order to take into account the contributions of all the peaks representing one protein [34, 67].

Most feature extraction methods, as described above, extract features from signals in the original space, *i.e.* peak intensities of mass spectra. In a few studies, features were extracted by projecting the signals from the original space onto another, usually lower-dimensional, space through linear transformations. Principal Component Analysis (PCA) has been widely used as a standard way for this purpose in many other data mining applications [141]. PCA identifies the orthogonal directions in which data vary maximally using the eigenvalue/eigenvector decomposition of the covariance matrix. Then the original signals are projected onto those directions, the number of which is usually smaller than the original dimension. The projections are called principal components and often used as features. Since only those directions that explain data variation maximally are selected in PCA, the projected data is of a lower dimension, but with a minimum loss of information. In one study, every m/z point was regarded as a dimension and PCA was applied to find principal components, which were used as features in clustering analysis [73]. The WT has been also employed not only to reduce

noise but also to extract features from mass spectra in a similar fashion as PCA is used [54, 77]. The WT also compresses data by projecting the original data onto *prespecified* orthogonal directions (wavelets). The coefficient of each wavelet becomes a feature in this case [54, 77]. Since the wavelets representing high frequency components are usually ignored, noise reduction is simultaneous accomplished with feature extraction. Both approaches are very sensitive to the choice of components (*i.e.,* principal eigenvectors in PCA or wavelets in the WT); therefore, it is important to determine criteria for selecting eigenvectors or wavelets prior to feature extraction. However, this is currently performed in an *ad hoc* manner. In addition, as compared with methods that select features in the original space, the features resulting from PCA or the WT are less interpretable because the features are extracted from the projected space. Thus, the inverse transformations are needed to reveal how features (m/z points) in the original space contribute to creating each feature in the projected space.

In feature extraction, a variety of peak detection and alignment algorithms are being developed and tested. The resolution and noise of mass spectrometry systems should be taken into account. For example, using the maximum peak of a peak group might lead to over/underestimation of relative abundance of a certain protein because it can be easily affected by noise. Likewise, peak alignment that only considers the mass error rate might deteriorate the sensitivity and specificity. It is possible that better diagnostic systems could be developed if more prior knowledge of mass spectrometry and the proteins present in blood was incorporated into the feature extraction process.

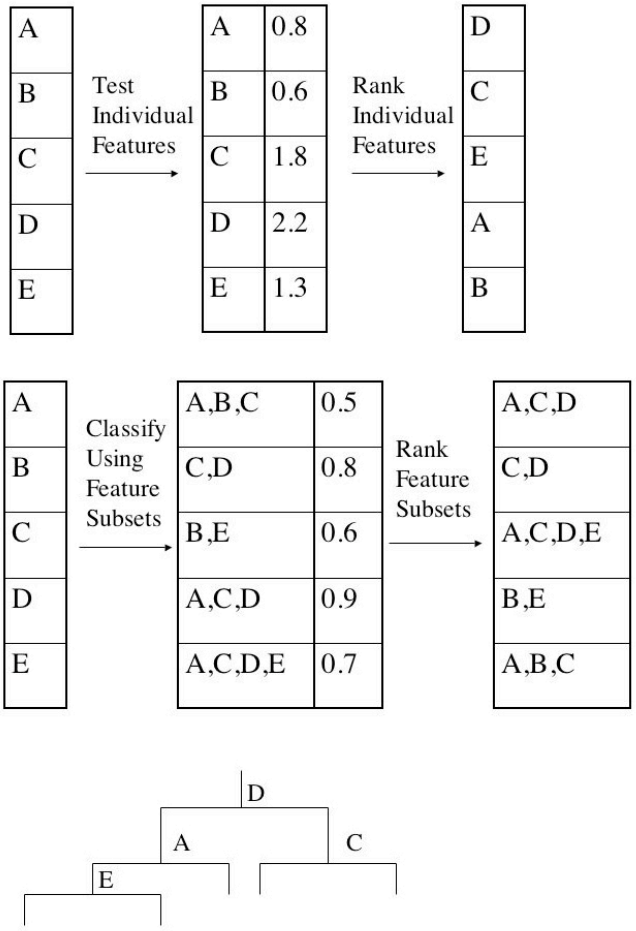

Figure 2.4: Feature selection methods are often categorized as filters (top panel), wrappers (middle panel), or embedded methods (bottom panel). A filter method evaluates and ranks individual features based on selection criteria (*e.g.*, t statistic). Then, a subset of features for classification is determined based on individual feature ranks. Wrappers assess the relevancy of a subset of features based on evaluation metrics of a classifier trained using that subset of features. Embedded methods implicitly perform feature selection as a part of the classifier training process (*e.g.*, decision tree).

2.3.3 Feature selection

The purpose of feature extraction is to produce a set of quantitative measures from a mass spectrum that could potentially be used for distinguishing spectra of normal and cancer samples. Typically, the feature extraction process results in a smaller set of

features (< 1000) than the number of (m/z, relative abundance) pairs that were in the original spectrum (≈15,000). However, in general, the number of features extracted is still much larger than the number of samples (< 500) in an experiment. This imbalance in the number of features and samples may increase the chances of misclassification due to overtraining and the usage of irrelevant or redundant features [41, 139, 142-144]. Also, a large number of features usually leads to an increase in the training time of classifiers. Moreover, from a biomedical perspective, it is important to find a moderate number of proteins that most contribute to correct classification such that these potential biomarkers can be identified and biochemically validated. Thus, it can be important to reduce the number of features from the set initially extracted. This process is referred to as feature selection.

Feature selection is defined as a series of actions to choose a subset of features that are relevant to correct classification based on specified evaluation and selection criteria [40, 139, 142, 144]. Feature selection methods are often categorized as filters, wrappers, or embedded methods (Figure 2.4). A filter method evaluates and ranks individual features based on selection criteria (*e.g.*, t statistic). Then, a subset of features for classification is determined based on individual feature ranks. Wrappers assess the relevancy of a subset of features based on evaluation metrics of a classifier trained using that subset of features. A search algorithm is used to explore the space of feature subsets and identify a high-performing subset of features. Cross-validation or bootstrap sampling is used in conjunction with wrapper methods since they can provide unbiased accuracy estimates of the classifier. Embedded methods implicitly perform feature selection as a part of the classifier training process.

Filters have been the most commonly used type of feature selection in prior studies of cancer classification using mass spectra. A variety of statistical tests have been

investigated to define selection criteria for the relevancy of individual features. The two-sample t test has been used in many studies [35, 37, 64, 79, 84, 85, 145]. A t test for two independent samples (cancer, normal) is performed on each feature across the training samples and features that show a statistically significant difference (*e.g.,* $p < 0.05$) in the group means are selected for use in training classifiers. Other studies have also used methods related to the t test for two independent samples. Li *et al*. define the distance between two sample groups, cancer and normal, as the absolute mean difference normalized by the root mean square of the variances of two sample groups [40]. This distance measure resembles the two-sample t test for independent samples with unequal variance. Zhu *et al*. calculated a reliable threshold for p value based on 1 D Gaussian random field considering the fact that multiple comparisons are made [35]. Other types of statistical tests such as the t test [82, 145], the one-way analysis of variance (ANOVA) [49, 68], the Wilcoxon signed rank test [38, 47, 65, 80] and the Mann-Whitney test [65] have also been used to rank features.

Some studies have tested the efficacy of relevancy measures on the basis of information theory and signal processing as filters. Information gain and relief-F [142, 146] are examples of measures used in information theory based filters [66]. The wavelet transform can also be used as a filter method for feature selection. In one study, features were assessed by comparing the wavelet coefficients of each feature between cancer and normal samples [77]. Receiver operating characteristics (ROC) analysis [147] has also been used to measure the relevancy of an individual feature. The area under the curve of each feature is calculated and it is used as the metric to rank features [51]. ROC analysis is discussed further in the evaluation section.

Using a single relevancy measure can lead to biased feature selection. Thus, combinations of methods have been investigated for feature selection [37, 70, 80]. A

feature is considered to be relevant when the feature receives high scores from multiple methods. This approach enables one to explore features from different perspectives and to make a more reliable decision regarding the selected subset of features.

Wrappers are different from filters in that classifier evaluation metrics are used rather than selection criteria for individual features and wrappers assess features in groups rather than individually. Filters employ selection criteria such as statistical tests to evaluate individual features, while wrappers use evaluation metrics of classifiers to estimate the discriminating power of a candidate subset of features [139, 142-144]. Moreover, while filters simply select a subset of features by choosing those that were highly ranked individually, wrappers iteratively optimize the subset selection using search algorithms such as genetic algorithms and stepwise selection methods [139, 142-144]. The wrapper approach typically has better performance than the filter approach since the search process in wrappers enables it to exclude redundant features when forming a subset of features [139, 142]. However, the filter approach does have the advantage that it is less computationally demanding than the wrapper approach [40, 142].

Several studies have investigated the efficacy of wrappers for feature selection in mass spectra. The combination of genetic algorithms [146] with classifiers is a popular use of wrappers in this field [36, 40, 71]. Several kinds of classifiers have been combined with genetic algorithms, including self-organizing maps [36, 42, 43, 56], support vector machines (SVM) [40], and simple distance based classifiers (*e.g.*, Mahalanobis distance) [54, 71, 84]. In other studies, stepwise feature selection methods (forward selection and backward elimination) [139] have been used instead [38, 67]. A wrapper that incorporates unified maximum separability analysis (UMSA) and bootstrap sampling has identified the best performing subset of features in three studies [50, 58, 83].

Embedded methods implicitly perform feature selection as a part of the classifier training process [139]. For example, decision trees estimate the contribution of individual features to correct classification in each iteration and grow the tree structure according to the estimation result. Therefore, when the training is over, the final subset of features is produced with the classifier [139, 146]. Feature selection using embedded methods for mass spectra will be further discussed in the next section on classification.

Feature selection can help to reduce running time and avoid overtraining if it succeeds in finding a subset of independent and discriminating features. Unfortunately, there is no guarantee that the feature selection process will improve the classification performance. Moreover, features selected as relevant for classification still need to be biologically validated in future studies. Efforts to identify the proteins corresponding to relevant features should follow feature selection and classification studies.

2.3.4 Classification

Machine learning is a branch of artificial intelligence that is concerned with design and application of algorithms that enable computers to learn from experience [146]. I interpret this definition broadly to include techniques that were developed from a statistical, rather than computer science perspective, such as linear discriminant analysis and regression.

There are three general types of machine learning algorithms: unsupervised, reinforcement, and supervised. In unsupervised learning, the computer attempts to identify natural groupings within a dataset based on criteria that define how "similar" items are and what makes a "good" group, but without being provided examples of the feature values of items and associated "correct" class membership. For this reason, unsupervised learning methods are also referred to as "clustering". Unsupervised learning

algorithms have not been used in many prior studies of cancer diagnosis from mass spectra. Some studies have explored self-organizing maps [36, 42, 43, 55, 56] and hierarchical clustering algorithms [60, 75, 79] in this field. In reinforcement learning, the computer is not provided with examples of the feature values of items and associated "correct" class membership, but is provided less specific feedback that indicates if the system is on the right track. I am unaware of any studies of mass spectra for cancer diagnosis that employ reinforcement learning methods. In supervised learning, the computer is provided with examples of the feature values of items and associated "correct" class membership. The goal of supervised learning is to develop a "classifier" that can predict the class membership from a set of pre-determined classes for an item based on a set of features that describe the item. Supervised learning methods have been used extensively in the investigation of cancer diagnosis from mass spectra. Prior studies have tested the performance of several supervised learning algorithms including artificial neural networks (ANN) [66, 76, 79, 89, 137] [148], k nearest neighbor (KNN) [35, 49, 66, 68, 137], logistic regression [37, 50, 58, 67, 83], decision trees [51, 59, 66, 67, 72, 77, 86, 137], linear or quadratic discriminant analysis (LDA/QDA) [39, 48, 49, 54, 67-69, 71], support vector machines (SVMs) [39, 40, 68, 137], matching pursuit (KMP) [63], logical analysis of data (LAD) [33], stepwise discriminant analysis [38], partial least square projection [73, 75], Naïve Bayes [66], rule induction [66], and ensemble algorithms (*e.g.* boosting, bagging, or random forest) combined with various base classifiers [32, 39, 52, 53]. Two evolving themes in the use of supervised learning in this field are the emphases on SVMs and ensemble methods.

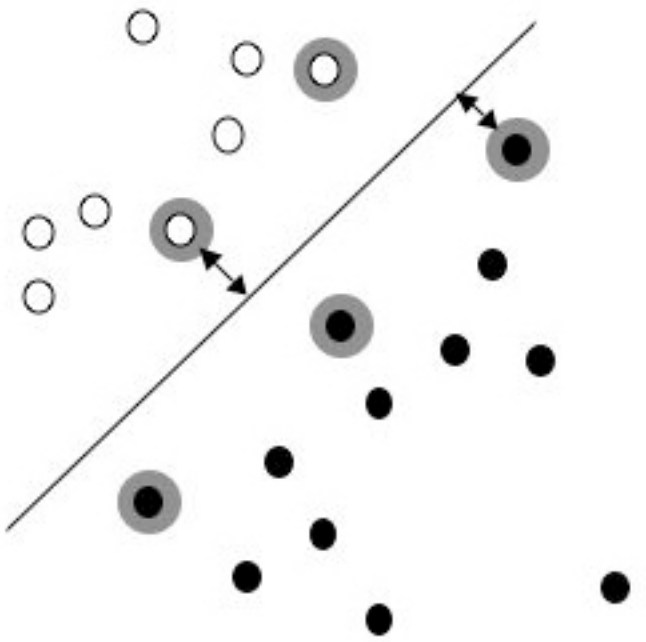

Figure 2.5: SVMs are a type of kernel learning methods, which project data from the current vector space to another vector space where linear learning algorithms can be applicable. SVMs guarantee the maximal margin between cancer and normal samples through global optimization of the decision boundary such that overtraining can easily be avoided (margin indicated by arrow). Since the decision boundary set by SVMs has the gradient that allows for the maximum-margin separation based on a few data samples closest to the decision boundary, which are called support vectors (highlighted with gray), SVMs implicitly reflect the contribution of each feature to successful classification and reduce the effect of irrelevant features by performing the dot product between the gradient and each sample.

SVM is a fairly new class of supervised machine learning methods that has generated considerable excitement (Figure 2.5). SVMs are a type of kernel learning methods, which project data from the current vector space to another vector space where linear learning algorithms can be applicable. The functions that project the data onto the new vector space, which usually has a higher dimension than the original, are called the kernel functions. Since an improper kernel function may worsen classification by

projecting data onto a space where linear separation is impossible, care must be taken for choosing a kernel function when using SVMs. Unfortunately, there are no guidelines for choosing the best kernel for a given data set. Prior knowledge of the characteristics of data may help with this process, but in practice selecting an optimal kernel remains a significant challenge. The most popularly used kernel functions are the polynomial, radial basis function, and sigmoid kernels.

After data projection into a linear space, SVMs guarantee the maximal margin between cancer and normal samples through global optimization of the decision boundary such that overtraining can easily be avoided [149, 150]. In the cases were the projected data is still not linearly separable, for example, when two classes overlap, a penalty is given to the objective function of optimization to trade the margin size and misclassification rate. A small penalty maximizes the margin size but increases the misclassification rate while a large one decreases the margin size but minimizes the misclassification rate [151].

SVMs can also be utilized without any data projection if the data are linearly separable in the current vector space. This method is usually called linear-SVMs. Since the decision boundary set by SVMs has the gradient that allows for the maximum-margin separation based on a few data samples closest to the decision boundary, which are called support vectors, SVMs implicitly reflect the contribution of each feature to successful classification and reduce the effect of irrelevant features by performing the dot product between the gradient and each sample. There is less need for an effective feature selection step when a classifier that is robust to irrelevant features is used. The robustness of SVMs to irrelevant and redundant features is especially valuable since mass spectra data sets typically have many more features than cases. Thus, SVMs exhibit several properties that are appealing in the analysis of mass spectra.

The complexity and subtlety of mass spectra patterns between cancer and normal samples may increase the chances of misclassification when a single classifier is used because a single classifier tends to cover patterns originating from only part of the sample space. Therefore, it would be beneficial if multiple classifiers could be trained in such a way that each of the classifiers covers a different part of the sample space and their classification results were integrated to produce the final classification.

Ensemble algorithms such as bagging, boosting, or random forests improve the classification performance by associating multiple base classifiers to work as a "committee" for decision-making [152, 153]. Any supervised learning algorithm can be used as a base classifier. Ensemble algorithms not only increase the classification accuracy, but also reduce the chances of overtraining since the committee avoids a biased decision by integrating the different predictions from the individual base classifiers.

Feature selection has been performed as an "embedded" part of the training process in many studies, especially when decision tree or SVM methods were used. Decision trees select the most discriminant features based on the information gain at each stage when growing the tree structure. As a result, a list of features that make the largest contributions to successful classification are obtained when classifier training is finished. In some studies of cancer classification using mass spectra, features selected implicitly by decision trees have been proposed as potential biomarkers [51, 53, 86]. SVMs also possess embedded feature selection mechanisms. As described in the earlier part of this section, the decision boundary includes the information of each feature's relevancy for successful classification. For example, in the case of the linear SVM, the absolute magnitude of coefficients of the decision boundary (a hyperplane) corresponds to the degree of relevancy of features. Prados *et al.* proposed a list of potential biomarkers using the internal feature selection function of a linear SVM [137].

The goal is to build *reliable* classifiers, which can classify unknown samples within a reasonably bounded error range. While the error of a classifier on the training set decreases as the training process proceeds, the error on the general population increases after a certain time point in the training process because the classifier becomes oversensitive to the patterns that exist only in the training set. This event is called as "overtraining." It is important to avoid overtraining by evaluating the classifier performance using an independent set of samples. In addition, it is impossible to find a classification algorithm superior to the others for all feature selection methods because every classification algorithm has its own learning bias [138, 146]. The performance of a classification algorithm can be varied by the choice of feature selection methods. For example, KNN is very sensitive to irrelevant and redundant features. However, in prior studies, the relationship between the chosen feature selection method and classification algorithm has not been thoroughly researched. It is necessary to identify the best pair of a feature selection method and classifier.

2.3.5 Evaluation

After a system is developed through the stages described in the previous sections, its performance must be carefully assessed. In this section I discuss two important issues in system evaluation. First, the quality of the data set used to develop the system will strongly influence its performance since systems for cancer diagnosis from mass spectra are inherently data-driven. Second, the system evaluation must be based on criteria that are clinically relevant and quantitative with clearly defined standards of interpretation.

The desired characteristics of the data are that they provide an accurate representation of the population to be tested and that there are sufficient data to allow for robust inference. There are many factors that can bias a sample such that it doesn't

correctly describe the population, *e.g.*, the choice of human subject inclusion/exclusion criteria, data entry errors, etc. This problem is complicated by the fact that disease cases typically must be present in the data set at a much higher proportion than the population prevalence in order to show the breadth of variability in the disease state with a limited overall sample size. The imbalance in the sizes of disease and healthy classes can make classifiers more sensitive to patterns originating from disease cases, resulting in more false positives in classification. If there are more healthy cases, the number of false negatives will increase because patterns from healthy cases will be relatively more emphasized. From this point of view, it is valuable to equalize the class sizes [154-156].

However, there would be difficulty in keeping the balance between disease and healthy sample sets as one attempts to increase the entire sample size for more robust and reliable inference because disease cases are usually more difficult to obtain than healthy ones. To the best of my knowledge, this issue has not yet been addressed in the arena of analyzing mass spectra. Over-sampling the minority class or under-sampling the majority class are common methods to resolve biased classification due to imbalanced data. The basic idea behind these techniques is to balance the sizes of two classes artificially. For example, over-sampling the minority class, *i.e.,* sampling with replacement, increases the size of the minority class up to that of the majority class. Similarly, under-sampling the majority class, *i.e.,* decimating samples, can reduce the size of the majority class up to that of the minority class. However, it should be noted that these two techniques must be carefully used because over-sampling a minority class may lead to overtraining to a specific pattern of the samples belonging to the minority class and under-sampling a majority class may lose some valuable patterns of majority class samples [154, 155]. Some studies have tried to resolve this issue by penalizing the error rates of the samples

of the minority class more, which prevents the classifier from sacrificing those samples of the minority class to decrease the overall error rates (*e.g.* 1-accuracy) [154-156].

Proper handling of mislabeled data samples is also an important issue for classifier training and evaluation. There are two approaches to contending with mislabeled data. One is to reduce the likelihood of its existence through experimental design and quality control. The second is to eliminate mislabeled data in post hoc fashion during the analysis. In practice, since even extremely rigorous experimental design and quality control may not be able to perfectly prevent the occurrence of mislabeled data, both approaches should be taken in order to alleviate the effects of mislabeled data on decision support systems [157-159].

In order to avoid mislabeled data through experimental design and quality control, one must consider the possible sources. For example, mislabeling can arise from data entry errors. To a large extent, this can be avoided through rigorous lab protocols. A more concerning source of mislabeled data is genuine confusion regarding the correct classification of a sample due to the error or limitations inherent to the diagnostic test used to establish truth or the absence of a test for truth. For example, a healthy sample may be mislabeled as positive based on a false-positive biopsy. This type of error can be avoided if samples are only included for study if they have undergone confirmatory testing (*e.g.*, repeated biopsy). On the other hand, a diseased sample can be mislabeled as healthy either because of a false-negative diagnostic test or because no diagnostic testing was performed (*e.g.*, an asymptomatic subject was presumed to be healthy). Given the limitations of existing diagnostic tests for detecting very early stage disease and the many reasons not to perform diagnostic tests on seemingly healthy individuals, this can be an important source of false-negative samples. The most common approach to avoid this problem is to only consider a healthy sample to be healthy after an appropriate duration

of disease-free follow-up time [160]. To the best of my knowledge, this issue has not been explicitly discussed in any reports of studies of cancer classification from mass spectrometry to date. Moreover, I am unaware of any studies that have demonstrated and analyzed the risk of system performance degradation due to mislabeled training/test data samples in the context of mass spectrometry analysis.

The machine learning literature can provide some guidance on post hoc methods for detecting mislabeled samples. Mislabeled samples may appear as outliers. Therefore, detecting mislabeled samples is closely related to detecting outliers. Some approaches for outlier detection have been developed. For example, simply analyzing the means and standard deviations of features with the confidence intervals of each feature can reveal outliers [159] because samples lying outside the confidence interval are highly probable to be outliers. Clustering algorithms also can be used to identify outliers [159, 161]. Presumably, samples belonging to the same class would be clustered together while outliers would behave as ones belonging to other classes. Note that this clustering should be performed prior to feature selection. Other studies have used multiple classifiers of different types to filter out outliers [162, 163]. The key idea is that the samples whose labels were consistent with the labels predicted by multiple classifiers were regarded as correct samples and that were not were regarded as outliers.

There is no theory to provide firm guidance on the sample sizes required to properly perform any of the stages of development of clinical decision support systems utilizing mass spectrometry of blood products. Sometimes, it is easy to identify in retrospect that a sample may have been too small, such as when an algorithm fails to converge or operates with unacceptably low performance. However, one needs to take care in devising evaluation strategies that help avoid the common and difficult problem of the system appearing to perform well on the data set used for development but proving

unsatisfactory when subjected to additional testing with more data. Fortunately, this danger can be reduced to a large degree through appropriate use of data partitioning and sampling schemes.

In general, three independent sets of samples are needed for the development and evaluation of a classification system [146]. One set is called the training set and used for training a classifier. During or after classifier training, the classifier should be pruned and adjusted to avoid possible overtraining using another, independent sample set, which is referred as the validation set. As described in the previous section on classifier training, the error on the validation set tends to increase after a certain time point while the error on the training set keeps decreasing as the training process continues. The time point at which the error on the validation set starts to increase is the point when training should conclude. The validation set is used to find the stopping point of training. After the classifier is developed using the training and validation sets, it must be evaluated with respect to the general population. The test set is used to estimate the true error of the classifier on the general population. It is also important to recognize that a mass spectrometry analysis is actually composed of a series of chemical/biochemical processes. Thus, within a data set samples must be randomized at each analytical step so as to avoid any possible bias due to batch processing because such bias could produce systematic patterns that interfere the "true" patterns originating from the pathological changes in the samples.

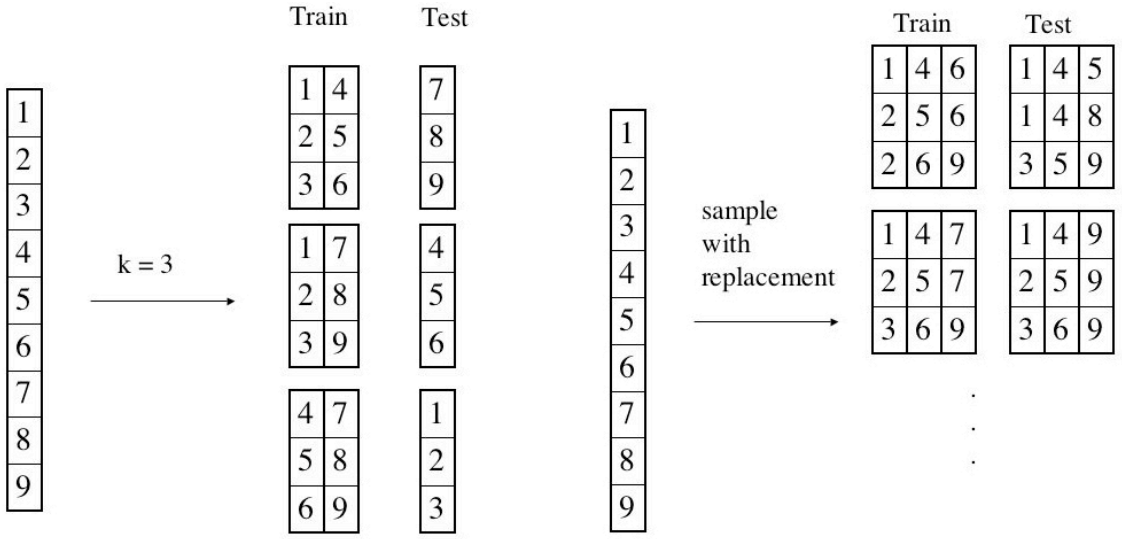

Figure 2.6: The left panel illustrates k-fold cross-validation and the right panel illustrates bootstrap sampling. In k-fold cross-validation, the data are split into k non-overlapping subsets or "folds" such that each sample is present in a single fold. The classifier is trained on k-1 of the folds and tested on the remaining fold. This process is repeated such that each fold is withheld once. Usually, the average of evaluation results (*e.g.*, accuracies) across the folds is taken as the estimate of the overall system performance. By comparison, a bootstrap set is created by random sampling of N cases with replacement from the original set of N cases. A classifier is trained on one such bootstrap set and tested on another. The process is repeated many times and the average of the evaluation results across the bootstrap sets is taken as the estimate of the system performance.

Typically, the same data (training set) are used in the procedures of preprocessing, feature extraction, feature selection, and classifier training. The use of separate sets for choosing algorithms and setting their parameters in each of these stages would provide greater protection against overtraining. Unfortunately, this is seldom plausible given realistic sample sizes. In fact, in most studies of cancer classification using mass spectra, the number of available samples is not even large enough to produce three independent sample sets. Even when three non-overlapping sets are used, they are typically partitioned from a single set and as such as are not truly "independent" sets. I am aware of very few

studies of cancer classification using mass spectra of human blood samples that have employed a truly independent test set (*e.g.*, test set was generated on a different day than the training set) [35].

The small number of cases necessitates the use of sampling techniques such as k-fold cross-validation, bootstrap sampling [164-166], or random partitioning (Figure 2.6) to estimate the generalization ability of the classifier. Sampling techniques are used to obtain estimates of classifier performance by judicious reuse of data. However, it should be noted that no sampling technique can perfectly address the question of how systematic and realistic variations in the data source (*e.g.,* variations in a single mass spectrometer over time or between two mass spectrometers) will impact the general classifier performance. A classification system must ultimately be evaluated using a large, independent data set.

In k-fold cross-validation, the data are split into k non-overlapping subsets or "folds" such that each sample is present in a single fold [164]. The classifier is trained on k-1 of the folds and tested on the remaining fold. This process is repeated such that each fold is withheld once. Usually, the average of the evaluation results (*e.g.,* accuracies) across the folds is taken as the estimate of the overall system performance [32, 33, 39, 42, 49, 53, 54, 58, 59, 62, 63, 66, 67, 69, 77, 79, 83, 86, 137]. When k is equal to the number of samples, this procedure is called leave-one-out cross-validation. In leave-one-out cross-validation, every sample is tested exactly one time and the overall system performance is estimated by simply gathering the individual sample validation results as if these test results came from a single classifier [33, 35, 40, 41, 69-73, 75, 77]. Note that the actual number of classifiers trained is equal to the value of k in the cross-validation.

It is important to remember that the performance estimates obtained by k-fold cross-validation are affected by the size of the training set and the number of folds. An

estimator is evaluated in terms of its bias, the extent to which the average system performance estimate is close to the true system performance in the population, and its variance, the extent to which the estimates spread around the average system performance estimate [167]. The estimate of the true system performance is more biased as the size of the training set decreases and has higher variance as the size of the testing set decreases [168, 169]. Therefore, a cross-validation using a larger value for k will result in an estimate with less bias, but higher variance relative to a cross-validation using a smaller value for k. Several excellent texts are available that discuss the trade-offs between bias and variance in classifier evaluation [138, 167-169].

The bootstrap sampling is another technique to estimate the true system performance with a limited number of samples [164, 165]. A bootstrap set is created by random sampling of N cases with replacement from the original set of N cases. A classifier is trained on one such bootstrap set and tested on another. The process is repeated many times and the average of the evaluation results across the bootstrap sets is taken as the estimate of the system performance [39, 42]. Note that each bootstrap set created for training results in a separate classifier. One study [39] employed 0.632+ bootstrap sampling, a modified version of bootstrap sampling, which can alleviate the bias in estimating the true system performance [168].

Random partitioning can be regarded as single or multiple 2 fold cross-validation. It is also similar to bootstrap sampling except that sampling is performed without replacement and less than N of N cases are selected. The training set is generated by randomly sampling a certain portion of data and the remaining samples of data are used as the test set [36-38, 42, 48, 51-56, 62, 67, 70, 80].

Commonly, cross-validation, bootstrap sampling, and random partitioning are used to estimate the system performance during the classifier training stage. However,

some studies have applied random partitioning to derive reliably discriminant features during feature selection [50, 58, 80, 83] based on the ranks of discriminant features that are earned on each sampled training data set. The features with consistently high ranks are selected for use. However, in practice, it is often impossible or very difficult for the *entire* design process to be performed in a cross-validation manner. As a consequence, several previous studies seem to have used the full data set prior to cross-validation for feature selection [39, 48, 68, 69, 72, 79, 83]. As was the case for estimating system performance, sampling techniques cannot overcome limitations that are inherent to the data set from which the samples are drawn. If the data set does not represent the underlying probability distribution of the population of interest, then even the most sophisticated feature selection methods based on sampling techniques will end up with an extremely "biased" subset of features [160].

It is critical that the system evaluation be based on criteria that are clinically relevant and be quantitative with clearly defined standards of interpretation. While classifiers typically attempt to optimize an evaluation function as part of the training process, it is important recognize that in general that function is not the most clinically relevant measure. For example, the mean-square error measure weights the two possible kinds of error equally while in most medical diagnostic tasks the costs, monetary and otherwise, of false-positives and false-negatives are not equal.

Accuracy, the fraction of the samples that the system correctly classifies, has been used in many mass spectrometry studies that employ a binary decision approach [32, 37, 39-41, 48, 49, 63, 66-68, 70-73, 75, 77, 87, 89, 137, 148]. However, there is a significant drawback to the accuracy metric in that it is dependant on the prevalence of disease in the data set. For example, if there are only 20 disease cases for every 80 normal cases, a system could achieve 80% accuracy by simply reporting all cases as normal. Thus, if the

prevalence is not 50%, the system accuracy can't be interpreted in isolation. The most clinically relevant measures for screening and diagnostic tests are sensitivity and specificity, regardless of whether the test involves a computational aid. Many studies of mass spectrometry for cancer classification have used these measures [33, 35-38, 43, 47, 48, 50-55, 58, 59, 62, 68, 69, 72, 79, 80, 83, 84, 86, 105, 137, 145].

Receiver Operating Characteristic (ROC) analysis can be used for diagnostic systems that provide a range of outputs rather than a binary classification. An ROC curve is a plot of the sensitivity vs. (1-specificity), or equivalently the true positive fraction vs. the false positive fraction, computed from the application of a series of thresholds to the system output (Figure 2.7). The advantage of ROC analysis is that it explicitly shows the tradeoffs in sensitivity and specificity that could be achieved with the same classification system. In essence, the choice of the decision threshold is delayed until a later time when more knowledge may be available on the costs associated with each type of error.

In general, ROC curves are concave and better system performance corresponds to more concave curves. A measure of the concaveness of ROC curves is the area under the curve (AUC). Hence, the AUC has been used as a measure of system performance in many studies [37, 40, 58, 69, 72, 80, 83, 86, 137, 148].

Evaluation metrics (*e.g.*, ROC AUC) are calculated based on a given data samples, yet it is the performance on the general population that matters. Therefore, there is a need to estimate the reliability of the system. For this purpose, some studies have randomly permuted the class labels of samples and compared the performance to that from using the actual class labels [49, 67, 68, 73]. As the difference between two becomes larger, the performance evaluation from the actual samples is taken as a more reliable indicator of how the system would perform on the general population.

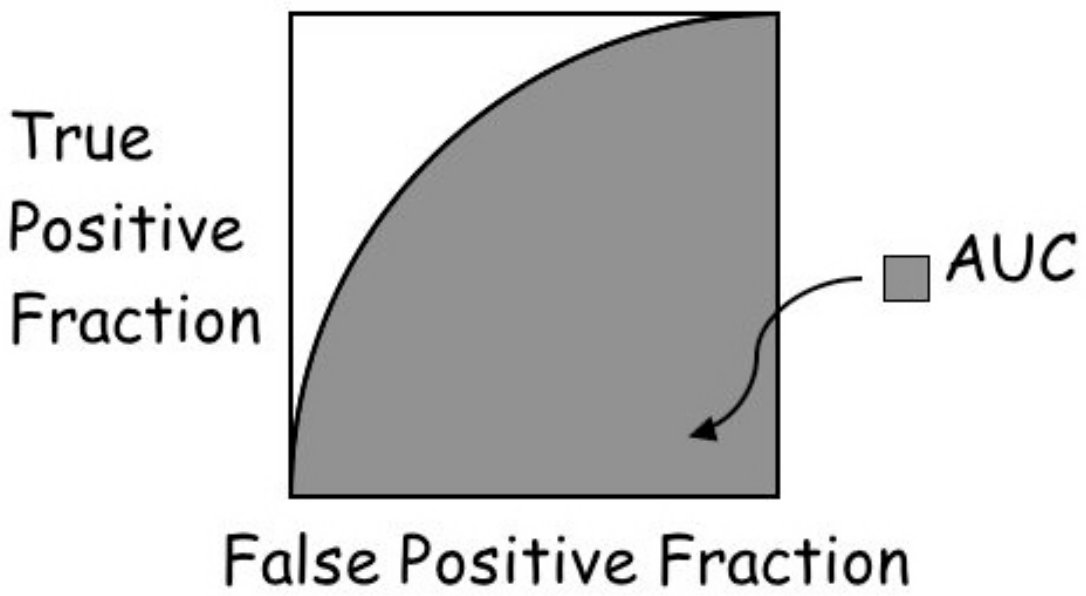

Figure 2.7: Receiver Operating Characteristic (ROC) analysis can be used to evaluate diagnostic systems that provide a range of outputs rather than a binary classification. An ROC curve is a plot of the sensitivity vs. (1-specificity), or equivalently the true positive fraction vs. the false positive fraction, computed from the application of a series of thresholds to the system output. A measure of the concaveness of ROC curves is the area under the curve (AUC).

When sampling techniques are used, care must be taken not to mistakenly tune classifier performance results on the "testing" portions. For example, several studies appear to have determined the threshold for calculating the sensitivity and specificity based on the "testing" portion of the data rather than the "training" portion [37, 40, 48, 58, 69, 72, 79, 80, 83]. These practices can partially undermine the protection against overtraining provided by those sample techniques.

The use of appropriate data sampling methods and relevant evaluation metrics can provide substantial reassurance that laboratory studies will contribute towards the goal of accurate and reliable clinical decision support systems. Of course, laboratory studies must

be followed by rigorous clinical testing. For example, studies of the way that the healthcare team does, or does not, incorporate the recommendations made by a system based on the mass spectrometry data are beyond the scope of this review. Ultimately, long-term, large clinical trials are required to establish the efficacy of any screening test to the level of a decrease in cause-specific mortality.

2.4 SUMMARY

An ideal screening method should be accurate, reliable, rapid, inexpensive, and minimally invasive. Proteomic profiling of blood samples using mass spectrometry has recently been proposed as a method that has the potential to meet these goals. However, there are key difficulties that must be addressed before clinical diagnostic tools can be developed based on this technology. Chief among these is to overcome the restrictions on reliability that have plagued early studies. To achieve accurate classification on a given set of samples is useless unless the classifier can also be generalized such that new, but similar, data can be accurately classified. A system for discriminating proteomic patterns of samples from healthy and ill people must be robust to the variability that will exist across people, mass spectrometers, sample collection protocols, days, etc.

This chapter reviews the literature on developing clinical decision support systems for cancer screening from proteomic patterns obtained by mass spectrometry of blood samples from a machine learning perspective. Prior studies are presented in an explicit machine learning framework consisting of five stages: preprocessing, feature extraction, feature selection, classifier training, and evaluation. The purpose of preprocessing is to reduce the influence of aspects of the data that are not expected to aid in the goal of discrimination between disease and healthy patterns and instead may make that classification task more difficult. In feature extraction, the aim is to reduce the

dimensionality of the data and increase the interpretability by defining numerical summary measures, often called "features". Following feature extraction, it is necessary to perform a feature selection step in which a subset of features that best enable discrimination between the two groups is identified. Given a set of spectra summarized by informative features and with corresponding truth (health status), a variety of classification algorithms can be trained. Finally, care must be taken in the choice of experimental design (*e.g.*, data sampling) and evaluation criteria in order to assess both accuracy and reliability (generalization).

It is apparent that the components of the framework that are most specific to the data type, mass spectra of blood samples, are preprocessing, feature extraction, and feature selection. I hypothesize that improvements in these components will yield the greatest increase in system reliability and that the approaches most likely to achieve those improvements will be based on explicit models of the data generation. While the objective of developing a clinical decision support system for cancer screening from proteomic patterns is ultimately data driven, I argue that this goal may not be achievable with reasonable sample sizes unless knowledge of the related biology, chemistry, and engineering is used to constrain the design process.

Chapter 3: ANALYSIS OF NOISE FROM INSTRUMENTATION IN MALDI TOF MASS SPECTROMETRY

3.1 INTRODUCTION

In recent years, matrix-assisted laser desorprtion/ionization time-of-flight (MALDI TOF) MS and its variants (*e.g.,* surface-enhanced laser desorption/ionization time-of-flight MS) analyzed with computational pattern recognition algorithms have attracted attention as tools for early diagnosis of cancer. The key role of MALDI TOF or SELDI TOF MS for early cancer diagnosis is to identify differences due to pathological changes between the mass spectra of diseased samples and those of controls so that pattern recognition algorithms can learn statistically dissimilar patterns. However, because most such pattern differences in mass spectra of samples such as plasma/serum are very subtle, noise can cause false positives or false negatives in peak detection by distorting the true shape of the mass spectrum. Thus, several studies have investigated methods for characterizing or reducing noise in order to improve the sensitivity of MS [35, 54, 63, 67, 68, 71, 73, 77, 103, 118, 119, 128-130, 132, 133, 170-173].

To date, most efforts for noise reduction, particularly in MALDI TOF MS, have focused on eliminating the baseline and reducing high frequency noise [35, 63, 68, 71, 73, 128, 129, 170, 172, 173]. The baseline is a monotonically decreasing bias in the mass spectrum that originates from matrix clusters formed during the ionizing process. To eliminate this baseline, it is heuristically estimated [63, 67, 68, 71, 103] and then subtracted from the original mass spectrum. For the baseline estimate, a local average or minimum intensity within a moving window [71], a piecewise linear regression line [67, 68], or the convex hull of the intensities [63] have often been used. On the other hand, high frequency noise appears in the mass spectrum as fast varying ripples or irregular

peaks. A number of factors such as electrical interference, random ion motions, statistical fluctuation in the detector gain, or chemical impurities may be involved with the occurrence of the high frequency noise. Heuristic approaches have been predominantly used to reduce high frequency noise. For example, moving average filters [63], Gaussian kernel filters [35, 132], principal component analysis (PCA) [172], and the wavelet transform (WT) [54, 77, 170] are common techniques for high frequency noise reduction. Andreev *et al.* obtained power spectral density estimates of the high frequency noise through non-parametric power spectral density estimation and designed a matched filter to reduce the noise adaptively [171]. Most manufacturers also provide noise reduction algorithms such as a moving average filter in their products; however, it is difficult to obtain optimal filtering results because the users must determine the filter parameters iteratively through experimentation or based on previous experience.

The noise reduction approaches introduced above have been established based on empirical insight rather than on rigorous statistical noise analysis; therefore, the parameters of these algorithms have been determined in an *ad hoc* manner. Few studies have investigated the noise sources in MS or attempted to model the noise by measuring its statistical characteristics. Anderle *et al.* attempted to represent the noise magnitude variance in liquid chromatographic MS (LC MS) as a combination of quadratic and linear models [133]. Similarly, Hastings *et al.* fitted the log transformed noise level to a sum of two normal distributions, and compared the performance of the average and median filters based on their noise model [173]. However, since these studies were not preformed from a stochastic signal processing perspective, they do not provide sufficient instinct on how noise varies with time and frequency. Malyarenko *et al.* developed a numerical baseline model using the phenomenon of exponentially decaying charge accumulation on the ion detector [129]. I proposed a noise model for MALDI TOF MS, where I

categorized noise into three types: noise from instrumentation, noise from random ion motions, and chemical noise [108, 130]. Then, I hypothesized that the observed noise is a result of multiplication and addition of these hidden components. Additionally, I reported the results of non-parametric power spectral density analysis on noise from instrumentation [130]. Similar efforts to reduce chemical noise were also made by some manufacturers. For example, Applied Biosystems Inc. developed an algorithm based on the Fourier transform and notch filtering to minimize the effect of chemical impurities on mass spectra [174]. They tried to identify periodic patterns of chemical noise in mass spectra using the Fourier transform, and to reduce signal deterioration by eliminating these periodic patterns using a notch filter. However, their approach does not seem to be strictly model-based in the sense that they did not build a model for chemical noise from the frequency representation.

These model-based studies represent an important advance over heuristic approaches. The lack of knowledge on statistical characteristics of the signal and noise in heuristic approaches may lead to the design of noise reduction algorithms or digital filters that deteriorate the true signal rather than restore it. However, more work needs to be done towards complete noise characterization. Prior studies may have oversimplified the noise sources or disregarded the importance of power spectral density analysis. For example, most noise analyses have not explicitly distinguished the subtypes of the high frequency noise; however, various electrical, physical, and chemical components of the mass spectrometer may generate subtypes of noise with different characteristics. Therefore, in order to elucidate the stochastic characteristics of noise in mass spectrometry, such individual noise components must be carefully separated and analyzed. In addition to noise subtype isolation and measurement, power spectral density estimation is also critical in noise characterization because this method can provide

guidance for digital filter design by showing the power distribution of noise over frequencies, which determines the magnitude and period of signal fluctuation due to noise in the mass spectrum.

As part of my effort for modeling noise in MALDI TOF mass spectrometry, I describe a method in which I have isolated noise from instrumentation occurring in the MALDI TOF mass spectrometer and obtained the signal model for this type of noise using parametric power spectral density estimation. By "noise from instrumentation", I mean the interference caused by electrical sources inside or near the mass spectrometer including thermal noise from the transimpedance amplifier, power supply and power line noise, and electrical interference from the ion accelerator pulse. My results show that noise from instrumentation is composed of $1/f$ noise and several prominent periodic components in addition to the common white spectrum due to thermal noise. However, it is encouraging that a simulation study based on the signal model of noise from instrumentation suggests that the errors due to only noise from instrumentation in mass spectra may not significantly hamper human observers or detection algorithms in finding meaningful patterns. The methodology could be applied to other types of noise or other types of mass spectrometry.

3.2 NOISE MODEL FOR MALDI TOF MASS SPECTROMETRY

I hypothesize that there exist three types of noise in MALDI TOF spectra: noise from instrumentation, noise from random ion motions, and chemical noise [108, 130]. Noise from instrumentation is generated by the random thermal motion of electrons in the circuits of the instrument or by electric/magnetic interference from nearby equipment. Noise from random ion motions originates from random ion motions and random fluctuations of the secondary electrons of the microchannel plate (MCP) ion detector.

Unlike noise from instrumentation, this type of noise is believed to be signal-dependent; thus, noise from random ion motions is hypothesized to be a multiplication of the signal intensity and a random variable. Chemical noise mainly originates from matrix clusters. Matrix is also ionized with the sample during the MALDI process and these matrix ions form clusters. The monotonically decreasing baseline is due to the presence of small clusters of matrix material, since the likelihood of cluster formation decreases with cluster size. Moreover, occasional sharp peaks in the higher mass ranges can result from unusually large clusters of matrix material. Among these three types of noise in MALDI TOF mass spectrometry, chemical noise is dominant. The proposed noise model is described as follows.

$$m(t) = R(t) \times x(t) + C(t) + I(t) \tag{3.1}$$

In Eq. 3.1, $x(t)$ and $m(t)$ represent the true signal and the measured signal at ion flying time t, respectively. As stated in Chapter 2, since an equal sampling interval is required for signal processing, time was used instead of mass to charge ratio in my noise analysis. $R(t)$, $C(t)$, and $I(t)$ symbolize noise from random ion motions, chemical noise, and noise from instrumentation, respectively. As you can see in equation (3.1), chemical noise $C(t)$ and noise from instrumentation $I(t)$ are added to the true signal $x(t)$ because these types of noise are independent from the true signal. Ion detection in mass spectrometry is based on counting statistics. The ion detector amplifies the number of ion arrivals at a given time bin utilizing the photomultiplier effect of the MCP. Therefore, this detection process includes two counting steps: ion arrival and amplification through the secondary electrons, which can be modeled using Poisson statistics. In Poisson statistics, the variance, which corresponds to noise power, is

proportional to the signal intensity $x(t)$; thus, noise from random ion motions $R(t)$ is described as a multiplication term to the true signal $x(t)$.

In this dissertation, noise from instrumentation $I(t)$ and chemical noise $C(t)$ have been analyzed using signal processing techniques. The analysis of noise from instrumentation is introduced in the later sections of this chapter, and that of chemical noise in Chapter 4.

3.3 FUNDAMENTAL THEORY

In general, a random process does not show regular patterns in the time domain like a sine wave because many signals of different frequencies and phases are added together. The power spectral density of a random process provides the power distribution of the signal with respect to frequencies. If there is a high value at a certain frequency in a power spectral density, the corresponding random process has a strong sine wave with that frequency in the time domain [175]. The simplest way of estimating the power spectral density of a random process is to calculate the absolute square of the Fourier transform of a given realization, which is referred to as the periodogram. Power spectral density estimation methods based on the periodogram are called nonparametric methods because these methods derive a power spectral density estimate from given realizations without any background information on the data source. However, nonparametric methods suffer from poor frequency resolution and spectral leakage effects due to the finite length of data. The lack of resolution in nonparametric estimation becomes more problematic when the sampling frequency is very high but the data length is relatively short. In this case, a non-parametric power spectral density estimate would provide power information on only a relatively small number of frequencies within a wide range of

frequencies [175]. Spectral leakage causes ripples in a power spectral density estimate, which make it difficult to identify true periodic components in the signal.

Parametric power spectral density analysis can overcome these drawbacks by estimating the parameters of a linear system under the assumption that the observed random signal is the output of the linear model when a random signal with a white frequency spectrum is given as input. Once a model is established, a high-resolution power spectral density estimate free from spectral leakage can be obtained since the power spectral density of the random signal is determined by the parameters of the linear system [175]. The differential equation between the input random signal and the observed signal in the time domain can be written as:

$$x(n) + a_1 x(n-1) + \cdots + a_p x(n-p) = w(n) + b_1 w(n-1) + \cdots + b_q (n-q) \tag{3.1}$$

In the above equation, $x(n)$ denotes the observed signal system at the n^{th} time index, and $w(n)$ the input random signal at the same time index. $H(f)$, the Fourier representation of the linear system, is defined as the ratio of $X(f)$ and $W(f)$, the Fourier representations of $x(n)$ and $w(n)$ and it is uniquely determined uniquely by $a_1,\ldots,a_p$ and $b_1,\ldots,b_q$. The power spectral density of the random signal $S_X(f)$ is obtained using the following equation:

$$S_X(f) = |H(f)|^2 S_w(f) \tag{3.2}$$

where $S_w(f)$ is the power spectral density of the input signal with a white spectrum [175].

Three different types of random processes can be generated using the linear model. When $b_1,\ldots,b_q = 0$, the process produced by the linear model is called an autoregressive (AR) process of order p. When $a_1,\ldots,a_p = 0$, the resulting process is called a moving average (MA) process of order q. Otherwise, the process is called an autoregressive-moving average (ARMA) process of order p and q. Generally, these three models could be exchanged if models of infinite order be allowed. However, among these three types, the AR model is most commonly used for power spectral estimation because it can show narrow frequency components more accurately than the others with simple linear equations for parameter estimation [175].

The Burg algorithm estimates the power spectral density using an AR model. The AR parameters are estimated by minimizing the forward and backward residuals of the model, which are defined as the error between the given random signal and their corresponding estimators at n and $n-p$ [175]. In general, power spectral density estimates obtained by the Burg algorithm have high frequency resolution [175], and are more unbiased and stable than other power spectral density estimation algorithms using an AR model such as the Yule-Walker algorithm and least square estimator [107].

Ideally, an infinite measurement of a random process is desired to develop a most accurate model; however, in reality, measurements have finite length due to practical limitations of instrumentation devices. For example, in MALDI TOF mass spectrometry, the maximum signal length is determined by the instrument according to a pre-defined limit on the maximum mass to charge ratio. In recognition of this common problem, de Waele and Broersen extended the Burg algorithm to obtain a more accurate model using multiple segments from a random process, than can be achieved using a single realization of the process [107]. Like the Burg algorithm, this algorithm also estimates the model parameters by minimizing the forward and backward residuals; however, the revised

algorithm attempts to minimize residuals from multiple segments simultaneously [107]. This extended Burg algorithm generates a more accurate model than parameter averaging methods, which develop a final model by averaging the parameters of the models derived from individual segments [107].

The model order must be carefully determined so that the model can represent the given segments well, while avoiding overfitting. In general, the residuals decrease as the model order increases, so the modeling process must be stopped at some point even though the residuals are still decreasing. In the Burg algorithm, the Akaike's information criterion (AIC) is employed to select the optimum model order [107]. The AIC is represented as the sum of the model order and the log residual of the model with respect to the given random process. The parameter estimation of the Burg algorithm stops when the AIC is minimized. When errors between the estimated model and true random process is normally distributed, the AIC is defined as the following equation

$$AIC(p) = \ln\left(RES(p)\right) + \frac{2p}{N} \tag{3.3}$$

where $RES(p)$ is the residual variance of the model of order p, N is the length of a given signal realization [107, 176]. In the Burg algorithm for multiple segments, the above definition of AIC is slightly modified so that it may reflect the fact that the variance of the estimated parameters becomes lower than when a single segment is used by a factor of S, which is the number of segments [107].

$$AIC_s(p) = \ln\left(RES(p)\right) + \frac{2p}{NS} \tag{3.4}$$

In this study, additional steps were taken to avoid overfitting. The Burg algorithm is prone to overfitting because it uses the same data to select the model order as are used to develop the model. Thus, in this study a portion of the data set was held out from the model development and used to select the final model. In this process, the final model was selected based on another metric, the Kullback-Leibler discrepancy (KLD). The KLD is a generalized error measure for two probabilistic distributions, $p(x)$ and $q(x)$ [177].

$$D(p \| q) = \int_x p(x) \log \frac{p(x)}{q(x)} \tag{3.5}$$

In this case, $p(x)$ represents the probabilistic distribution estimate of the model from the Burg algorithm, and $p(x)$ the probabilistic distribution of the held-out set. In fact, the AIC is an estimate of the KLD that is specialized for measuring the distance between a set of realizations of a random process and a model developed based on them [178, 179]. However, in general, the AIC may not be appropriate for estimating the distance from a model to another independent set [179]; thus, the KLD was adopted for selecting the final model using the held-out set.

3.4 MATERIALS AND METHODS

MALDI TOF mass spectra were measured from a blank plate to obtain noise from instrumentation. This type of noise is generated by electric circuits (*e.g.*, the transimpedance amplifier, power supply and power line, and the ion accelerator pulse) in the instrument and electric/magnetic interferences from nearby equipment. Since no actual ion particle detection is performed in the experiments, noise from instrumentation

does not include the noise caused by the ion detector. Since the gold coating of the plate can cause chemical noise if the laser hits it, I ensured that the laser was not directly illuminating the plate by installing a physical barrier between them. A total of six data sets were created using three MALDI TOF machines of two types to investigate how the power spectral density of noise from instrumentation varies with machine type, location, and time. Data were collected on October 7th, 2005 and October 17th, 2005 using two Voyager Biospectrometry instruments (Applied Biosystems, Framingham, MA) located in two separate proteomics core facilities of The University of Texas at Austin (UT). The acceleration voltage of the mass analyzer was set to 28,125 V. Each spectrum was the average of 256 individual scans and had 262,144 data points with a bin size of 10 ns (sampling rate). Each UT data set consisted of 20 mass spectra. Averaging multiple scans to obtain a mass spectrum has been traditionally accepted to reduce the randomness that may occur in data acquisition, which can be considered as an elementary noise reduction scheme. Therefore, I investigated the potential effects of noise from instrumentation on mass spectra by deriving an AR model based on the average of individual scans. It should be noted that the average of individual scans is still a random process, so statistics like the PSD can be derived from it. Data were also collected on November 4th, 2005 and November 21st, 2005 using a third machine, a Voyager STR MALDI TOF instrument (Applied Biosystems, Framingham, MA), located at the Moffitt Cancer Center (MCC). The acceleration voltage of the mass analyzer was set to 25,000 V. Each mass spectrum was the average of 250 scans and had 233,889 data points with a bin size of 10 ns. Each MCC data set consisted of 20 mass spectra. In each data set, 10 mass spectra were randomly selected and held out as a validation group to determine the optimal model order and the remaining 10 mass spectra were used to develop a linear model for noise from instrumentation.

The Burg algorithm for multiple segments was applied to the training portion of each of the six data sets to obtain an AR model for noise from instrumentation for each of the machines. Because the DC offset of mass spectra introduces bias in the model parameters, the DC offset must be estimated and subtracted [107]. In my study, the means of individual mass spectra were used as the estimate of the DC offset. The Burg algorithm for segments was implemented by de Waele and Broersen [107] using MATLAB® (TheMathworks, Natick, MA), and their toolbox is publicly available (http://www.mathworks.com). This MATLAB® implementation allows the user to limit the maximum model order to control the complexity of the model. The Burg algorithm for segments was used to develop a model on the training portion of the data. The algorithm uses AIC to select the optimal model order, on the training data, up to the specified maximum model order. The entire process was repeated several times with the maximal model order parameter varied from 100 to 10,000. The final model was selected from among this set of possible models using the validation set. The average KLD between each model and the held-out mass spectra was calculated and the model with the smallest average KLD was selected as the optimal model for the data set.

Once the final models for the data sets were determined, the power spectral densities of the models were obtained using a Fourier transform from the model parameters. A sharp peak of the power spectral density at a certain frequency means that a strong sine wave with the frequency exists in the noise. However, in order to fully understand how noise from instrumentation affects mass spectra, a true signal without noise (*e.g.*, mass spectrum free from noise) would also be needed. Since this cannot be obtained in general, a simulation was performed in my study in order to reveal the effect of noise from instrumentation on MALDI TOF mass spectra.

The potential effect of noise from instrumentation was investigated by adding simulated noise to simulated noise-free MALDI TOF mass spectra. Noise from instrumentation was simulated based on data generation methods proposed by Broersen and de Waele [180], which can generate a random process given an AR model obtained from the Burg algorithm. Because the noise generator produces a standard stationary random signal with zero-mean and unit-standard deviation, the simulated noise was compensated to have the mean and standard deviation estimated from real mass spectra of noise from instrumentation. Noise-free MALDI TOF mass spectra were simulated using the MALDI TOF simulation model developed by Coombes *et al.* [181], which I translated from S-PLUS® (Insightful Corp., Seatle, WA) to MATLAB®. Coombes *et al.*'s MALDI TOF model includes several key aspects of the MALDI TOF process such as peak broadening due to the distribution of isotopes and initial ion velocities. Generally, 100s-1,000s molecules are ionized per laser shot with initial velocities whose mean and standard deviation are 350 m/s and 50 m/s respectively during the MALDI TOF process [182, 183]. In my simulation, it was assumed that 1,000 molecules ($\approx 1.7 \times 10^{-21}$ moles) are ionized in each laser shot. Microchannel plate (MCP) detectors, commonly used in MALDI TOF, amplify the signal for detected ions by a factor of 10^2-10^4 [184]. Generally, TOF mass spectrometers employ the chevron MCP as a detector, which provides a gain of about 10^6-10^7 per ion collision [185]. Since the specifications of the transimpedance amplifier after the detector are not publicly available, my simulation assumes a total gain of 10^7 in ion detection and that the MCP generates no additional noise (*e.g.,* shot noise in the detector). A total of 57 proteins contained in human plasma were simulated. The number of proteins molecules ionized by the MALDI process was calculated based on the relative concentration ratios of these proteins in human plasma [126]. Each simulated mass spectrum was assumed to be externally calibrated using six calibrants (m/z =

175.2, 1060, 5734, 12360.5, 16951.5, 66430: arginine, bradykinin, bovine insulin, cytochrome C, myoglobin, bovine serum albumin) using the least square error method.

3.5 EXPERIMENTAL RESULTS

In a plot of the power spectral density, the x-axis represents the frequency (linear scale) and the y-axis represents the normalized power of each periodic component in noise (logarithmic scale). In general, a mass spectrum shows the relative abundances of protein/peptide species given a sample, which are actually the digitized values of the output voltage from the transimpedance amplifier connected to the ion detector; however, since the units of those values are not provided by the manufacturer, the unit of PSD cannot be specified in this chapter. The power spectral density was normalized with respect to the power gain between the input, in this case a white Gaussian random signal with an unit variance, and the output of the linear signal model established by the Burg algorithm for segments. The power spectral densities for spectra collected on the same machine on different days are similar (*e.g.*, compare Figure 3.1 A and B). Thus, the power spectral density of noise from instrumentation remains stable over the time scale of this study. It was observed that the noise power at 0 Hz is non-zero and monotonically decreases until about 5 kHz. This power component at 0 Hz may be caused by the bias between the estimated DC offset of mass spectra and the true value. This bias may slightly affect the model parameters, resulting in the DC power component in the power spectral density [107].

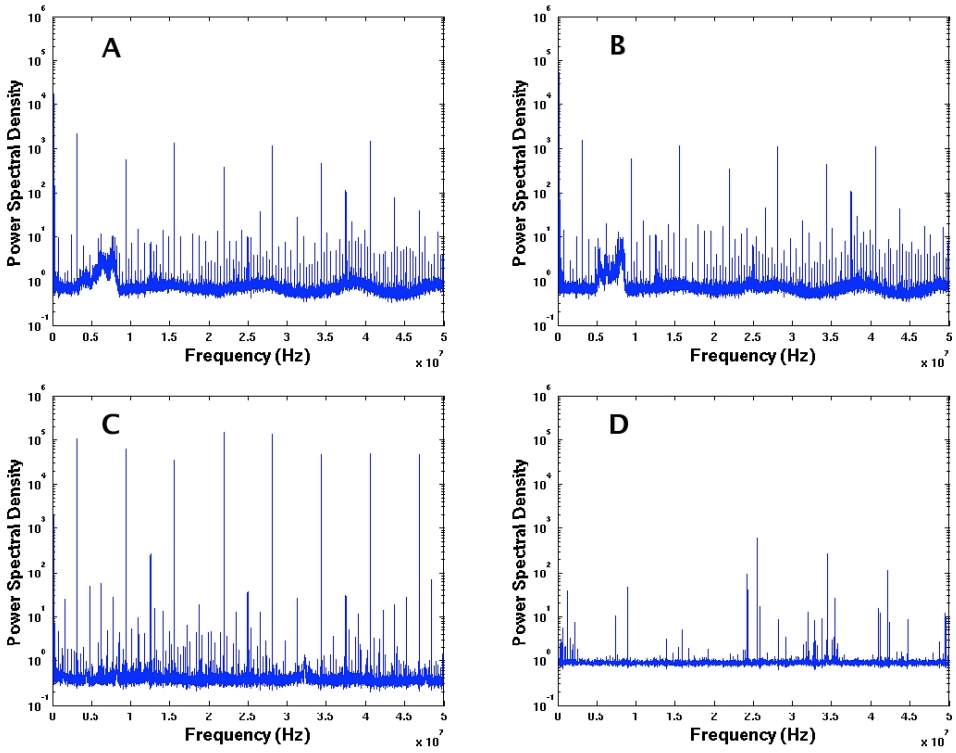

Figure 3.1: Power spectral densities of the AR models obtained from (A) SetA_UT1, (B) SetB_UT1, (C) SetA_UT2, and (D) SetA_MCC. When comparing (A) and (B), the frequency characteristics of noise from instrumentation in the same MALDI TOF instrument does not vary over dates of collection. Two MALDI TOF instrument of the older model type (Voyager Biospectrometry) show similar power spectral densities ((B) and (C)) containing prominent harmonics and more periodic components. In comparison, the instrument of the newer model type (Voyager STR) shows no noticeable harmonics and fewer periodic components in its power spectral density ((D)).

Nonetheless, this monotonically decreasing component may not be completely explained by the bias in the model parameter. One possible hypothesis is that this noise component may originate from $1/f$ noise of the MALDI TOF instrument. This type of noise is mainly introduced by a fluctuation of the mobility of the free charge carriers in an electronic device, and it is characterized by the inverse relationship between the frequency and the power spectrum [186].

Many peaks are observed in the power spectral densities of the data from UT, which suggests that mass spectra from those instruments may be affected by electric or magnetic interferences in addition to thermal noise. Harmonics that begin at 3.125 MHz and continue at an interval of 6.25 MHz until 40.625 MHz are present in the power spectral densities of the UT instruments, which are identical models located in separate facilities. The fact that the harmonics are observed in both devices at UT implies that the source of this interference is within the mass spectrometer (compare Figure 3.1 A and C). On the other hand, there are non-harmonic periodic components present in the power spectral density for one of the UT instruments but not the other (compare Figure 3.1 A and C). The absence of these periodic components in the UT2 power spectral density suggests that external sources generating electric or magnetic interference ranging from 5 MHz to 10 MHz may exist near the UT1 MALDI TOF instrument, but not near UT2 since these instruments are the same machine type, but located in different facilities. In principle, this hypothesis could be tested by systematically turning off all other instruments in the facility and re-analyzing the mass spectra of noise from instrumentation.

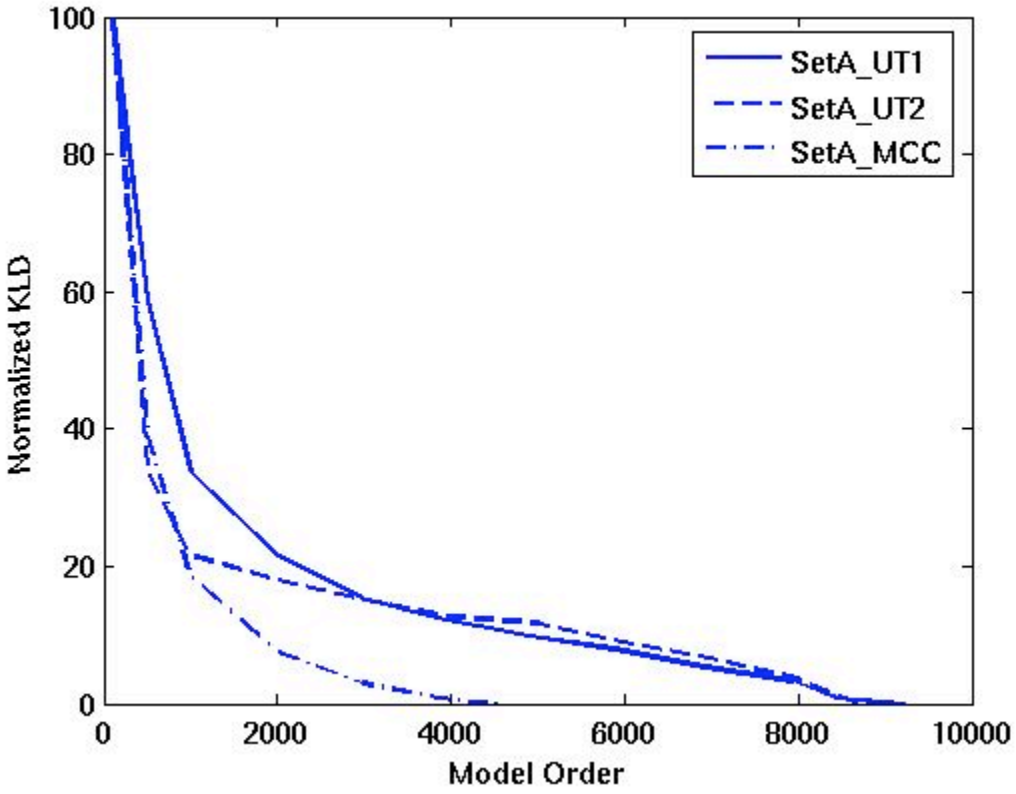

Figure 3.2: Normalized KLDs of the AR models with respect to the validation mass spectra. The KLD of each AR model is normalized with respect to its maximum and minimum values, and then multiplied by 100. The solid line is the KLD of SetA_UT1, the dashed line is that of SetA_UT2, and the dash dot line is that of SetA_MCC. The optimal model order of each model is decided at the point where its KLD stops decreasing.

The power spectral densities of different models of MALDI TOF instruments were also compared (Figure 3.1). Unlike the power spectral densities of the UT instruments, the power spectral densities of the spectra from the MCC machine do not have regular patterns like harmonics. Moreover, fewer periodic components were observed in the MCC power spectral densities than in those of the instruments at UT Austin (compare Figure 3.1 A and D). The average KLD of AR models with respect to the validation mass spectra provides additional evidence (Figure 3.2). For SetA_UT1, the

average KLD decreases as the model order is increased up to about 9,000 and then plateaus; thus, the optimal model order is the maximum order of 9,000. Similarly, the optimal AR model order for SetA_UT2 is approximately 8,500. However, the KLD of the model for SetA_MCC plateaus at about 4,500, which suggests that the power spectral density of SetA_MCC may contain fewer periodic components than those of SetA_UT1 and SetA_UT2 since each term in the AR model represents a periodic component with a specific frequency. The power spectral density and model order optimization analyses imply that the newer MALDI TOF instrument (Voyager STR, Applied Biosystems, Framingham, MA) at MCC may employ more effective electro-magnetic shielding schemes than the earlier model.

The power spectral density of noise is extremely useful when designing digital filters because the power spectral density indicates which periodic components are dominant in signal deterioration, and thus should be removed. However, it is difficult to determine how noise from instrumentation affects mass spectra by looking at only their power spectral density. Thus, the impact of noise from instrumentation was investigated by adding noise simulated based on the noise model to simulated noise-free MALDI TOF mass spectra. Figure 3.3 A presents the full view of the simulated mass spectrum without any types of noise that exists in MALDI TOF mass spectrometry. If the DC offsets in the simulated noisy spectra are ignored, there is little change in peak shapes relative to the simulated noise-free spectra (*e.g.,* Figure 3.3). This is consistent with the fact that the average root-mean-square (RMS) magnitude of noise from instrumentation ranges only from 6 to 11 (Table 3.1), which is negligible compared to the height of peaks in a mass spectrum.

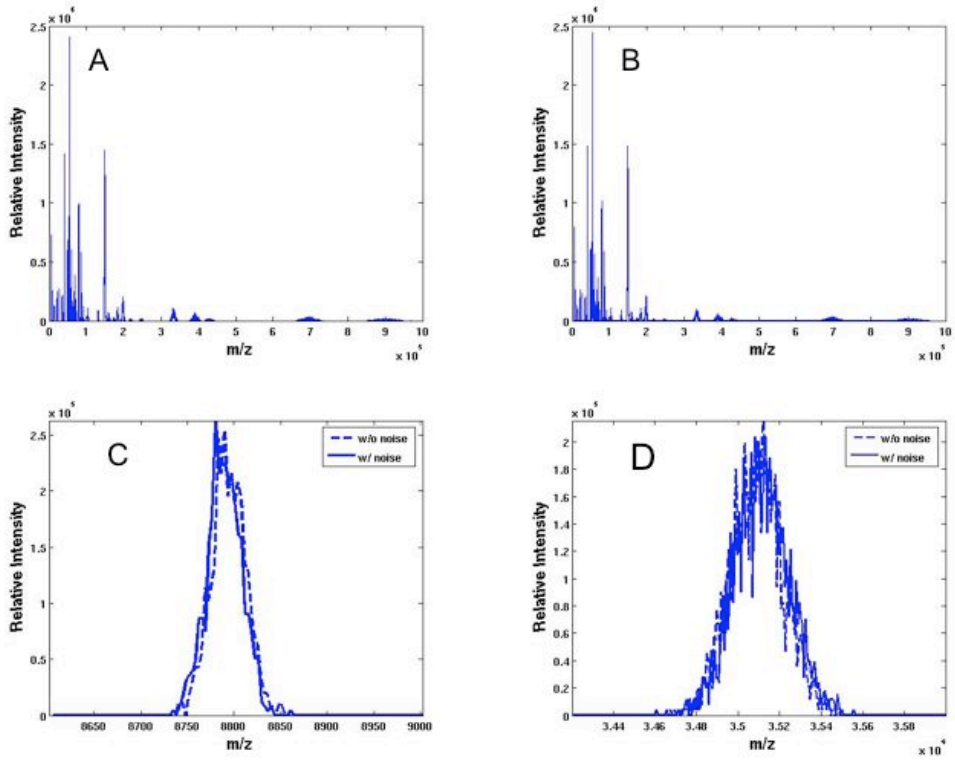

Figure 3.3: Simulated human plasma mass spectra. It is assumed that about 1,000 molecules are ionized every laser illumination, and the gain of the ion detector is 10^7. (A) The entire view of the mass spectrum without noise from instrumentation. (B) The entire view of the mass spectrum with noise from instrumentation. (C) A zoomed view of a MALDI mass spectrum showing a peak near m/z 8,800. (D) A zoomed view of mass spectrum near 35,000 Da. In (C), and (D), the solid lines represent mass spectra with noise, and the dashed lines mass spectra without noise. In (D), the peak with noise from instrumentation is not clearly distinguished from that without noise from instrumentation.

Table 3.1: The average DC offset and average RMS magnitude of mass spectra in each data set in relative intensity. As can be seen in this table, these statistics are consistent over time, but vary across the instruments. The potential effect of noise from instrumentation was investigated by adding simulated noise to simulated noise-free MALDI TOF mass spectra. These DC offsets and RMS magnitudes are needed in generating simulation noise using the models obtained my parametric power spectral density analysis.

Data Set	DC offset	RMS Magnitude
SetA_UT1	255.0	6.9
SetB_UT1	281.4	6.8
SetA_UT2	844.5	11.9
SetB_UT2	905.9	11.2
SetA_MCC	1425.4	6.8
SetB_MCC	1523.5	5.4

As can be seen in Table 3.1, each of the instruments has seemingly consistent average DC offsets and average RMS magnitudes over time. The quality of mass spectra is affected by the average RMS magnitude of noise from instrumentation. That is, if the average RMS magnitude of noise from instrumentation of an instrument is large, mass spectra obtained by the instrument are significantly contaminated by noise from instrumentation. However, as can be seen in Table 3.1, the average RMS magnitude of noise from instrumentation ranges from about 6 to 12, which implies that the potential effect of noise from instrumentation is negligible.

3.6 DISCUSSIONS AND CONCLUSION

The power spectral density reveals how the power of the periodic components hidden in the noise is distributed with respect to frequencies given a random process, and thus helps in developing filtering strategies for noise reduction. In my study, noise from instrumentation was separated from other types noise in MALDI TOF MS, and its power spectral density was estimated using the Burg algorithm for multiple segments, which develops an AR model for the noise by minimizing the residuals between the model and multiple observed noise segments simultaneously. The Burg method for segments provides much less biased models than other methods such as parameter averaging methods when multiple signal segments from the same source are available for parameter estimation [107]. Thus, this algorithm is well suited for the purpose of estimating the power spectral density of a random process with a finite length, but with multiple realizations available, such as is the case for noise from instrumentation. To see the variation of the power spectral density with the instrument type, location, and date of collection, six data sets of noise from instrumentation were measured from three different MALDI TOF instruments. The power spectral density does not vary much over the time scale studied, but it varies with the instrument type and location. The comparison of the power spectral densities from two identical instruments located in the different facilities suggests that both internal and external electric or magnetic interference sources affect the mass spectra. Therefore, shielding should be carefully considered to avoid signal deterioration due to the interference from nearby equipment. Interestingly, in the comparison between the newer and older models, fewer periodic components are seen in the power spectral densities of the newer instrument than in those of the older ones. This is probably a consequence of more advanced instrumentation design of the newer model that provides better shielding to the internal or external interference.

The potential effect of noise from instrumentation was investigated through a simulation study. The simulation suggests that noise from instrumentation may not significantly impact the interpretation of mass spectra. In fact, the RMS magnitude of noise is almost negligible in the high mass region when it compared to the randomness of the peak shapes due to ions' random initial velocities.

In conclusion, this chapter presents a systemic methodology for modeling noise from instrumentation in MALDI TOF MS on the basis of parametric power spectral density estimation using multiple realizations. My study opens a way of isolating a noise component, and measuring its stochastic features, which are critical in designing filters for signal manipulation often needed for MS applications like biomarker identification. In addition, this methodology will also benefit system designers of mass spectrometers as well by providing reliable spectral information on noise, letting them developing better shielding strategies for potential signal interference. For example, in my study, the power spectral densities of the mass spectrometers of the earlier model indicate that more shielding should be considered to avoid the periodic interference for a higher signal quality although the overall impact of noise from instrumentation was assessed to be low according to my simulation study. Isolating individual subtypes of noise and performing stochastic modeling of them will provide an important perspective on how to suppress signal deterioration due to the noise effectively by showing the power distribution over frequencies. Furthermore, such noise analysis can also be extended to other types of instrumentation like ESI MS once the types of noise in the instrumentation are identified and isolated. Hence, this technique is expected to benefit noise reduction studies for other types of MS instrumentation as well.

Chapter 4: ANALYSIS OF CHEMICAL NOISE IN MALDI TOF MASS SPECTROMETRY

4.1 INTRODUCTION

In the literature of mass spectrometry, "chemical noise" refers to many types of interferences due to chemical impurities in the sample. This definition covers a very wide range of sources of potential interferences in mass spectrometry, which sometimes leads to misunderstanding and confusion. Moreover, since chemical impurities in the sample depend on the type of mass spectrometry (*e.g.*, MALDI TOF or ESI) and analytes, the patterns of chemical noise are highly variable. Thus, it is necessary to first define chemical noise of MALDI TOF more precisely in the context of my noise analysis study.

In my study, chemical noise is defined as the unwanted interferences due to matrix in mass spectra. Matrix is organic material with a small molecular weight. In MALDI mass spectrometry, sinapinic acid or α-cyano-4-hydroxycinnamic acid is often used. It is mixed in large molar excess compared to the proteins and peptides. The primary role of the matrix material is to absorb the laser energy and transfer it to the sample, helping the molecules to be ionized without damage due to high laser power. The matrix is also used to physically separate the various components of the sample, preventing aggregation or precipitation [114]. However, the interaction of matrix material with itself, *i.e.*, clustering, introduces noise into the measured signal. This noise from chemical impurities mainly manifests as a monotonically decreasing baseline and a high signal variance in the low mass to charge ratio region.

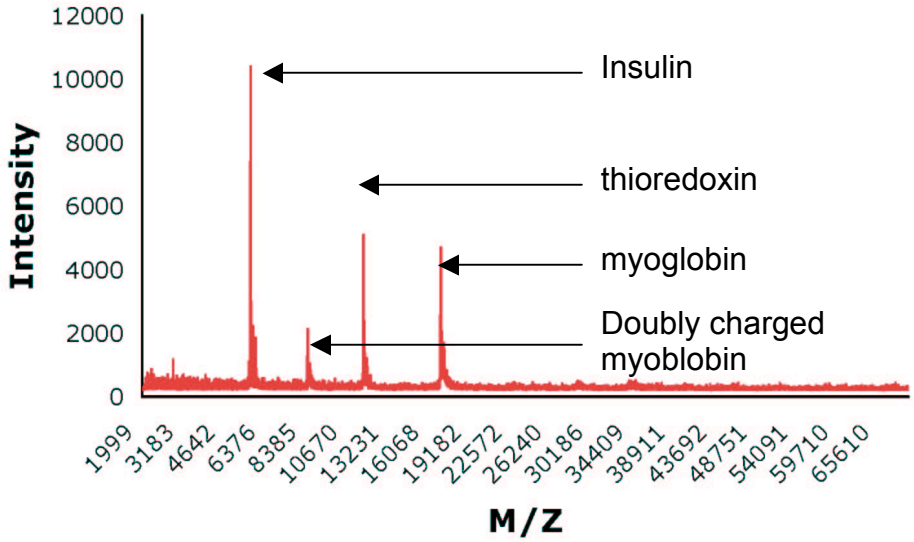

Figure 4.1: An example MALDI TOF mass spectrum of typical calibration proteins: insulin, thioredoxin, and myoblobin. The low mass region is significantly affected by chemical noise.

Figure 4.1 shows an example MALDI TOF mass spectrum of typical calibration proteins: insulin (5,808 Da), thioredoxin (12,000 Da), and myoglobin (16,952 Da). The second peak represents doubly charged molecules of myoblogin (≈ 8,000). As can be seen in Figure 4.1, a monotonically decreasing baseline due to matrix material extends up to nearly 60,000 Da, and the low mass region is corrupted by noise with high variance. It is observed that the variance of the noise also decreases with the baseline.

Many methods for reducing chemical noise in mass spectra have been developed and tested. Prior studies of noise reduction for mass spectrometry were extensively reviewed in Chapter 2 (Preprocessing) and Chapter 3 (Introduction). Since every method has its own advantages and disadvantages, it is difficult to conclude if one method is

better than the others. As discussed in Chapter 3, the current approaches to chemical noise reduction in mass spectrometry can be summarized in two categories: heuristic methods and model-based methods. Heuristic approaches attempt to remove the baseline or high-frequency noise using algorithms whose parameters are determined in an *ad hoc* manner while model-based approaches build a mathematical model and perform noise reduction based on the model.

However, conventional noise reduction schemes are not effective for denoising chemical noise because the characteristics of chemical noise are different from those assumed in the model (*e.g.,* Gaussian white noise). Because chemical noise originates from real chemical impurities, its characteristics are very similar to those of the mass spectrometry signal from the analyte, which makes it difficult to distinguish chemical noise from the real signal. Moreover, chemical noise is non-white and non-stationary. From a signal processing point of view, chemical noise is a mixture of frequency components of different powers and the power of each frequency component varies over time. Thus, conventional noise reduction schemes based on Fourier analysis fail to eliminate chemical noise.

The wavelet transform (WT) is a signal analysis method that decomposes a given 1D or 2D signal into frequency sub-bands over time. While the Fourier transform only shows the frequency distribution of a given signal, the WT describes the change of the distribution over time. This is why the WT is used for analysis of non-stationary signals.

Denoising via the WT is mostly done by thresholding. After decomposing the given signal using the WT, a threshold is set for each of the frequency sub-bands and the wavelet coefficients below the threshold are removed. Then the denoised signal is reconstructed by the inverse wavelet transform (IWT) of the shrunk wavelet coefficients [187, 188]. One of the benefits of wavelet denoising is that it can selectively reduce high

frequency noise while preserving the high frequency structure of the signal because the WT can provide the time location as well as the magnitude of a certain frequency component. In contrast, noise reduction through Fourier analysis simply rejects high-frequency components above a cut-off frequency, resulting in the loss of high frequency components of the signal [134].

There have been several studies of application of the WT to noise reduction in mass spectrometry [54, 77, 135, 170, 189]. Barclay and Bonner extensively studied the efficacy of the discrete WT (DWT) as a denoising method for mass spectrometry as compared to digital filtering (Fourier transform) and a heuristic smoothing algorithm (Savitzky-Golay) [135]. They applied these three methods to synthetic Gaussian and triangular mass spectra and real experimental ESI mass spectra. Barclay and Bonner observed that denoising using the DWT effectively removes the high frequency noise but keeps the narrow peak shapes while digital filtering based on the Fourier transform lost most of the narrow peaks [135]. Their study shows that the WT does not blindly smear out all the high frequency components above a certain cut-off frequency, but instead it successfully discerns the signal and noise that have similar frequencies. Qu *et al.* also applied the DWT to analyzing SELDI TOF mass spectra of prostate cancer serum samples [54]. They attempted to compress the mass spectra by removing the noise components through thresholding and then reconstructing the thresholded wavelet coefficients. They concluded that data compression based on the DWT can retain patterns useful for disease detection [54]. Denoising through the DWT can be done very quickly through the fast wavelet transform algorithm [190]. However, since the wavelet coefficients obtained by the DWT are not shift-invariant, the denoised signal may suffer from some artifacts near discontinuities (pseudo-Gibson phenomenon) [191]. Another problem is that the denoising results highly depend on the shift of the input signal [170].

To resolve these issues, Coombes *et al.* adopted the discrete stationary wavelet transform (SWT) to denoise SELDI TOF mass spectra [170]. The SWT is often referred to as the undecimated discrete wavelet transform (UDWT) in the literature. This variant of the WT algorithm produces shift-invariant wavelet coefficients of a given signal by averaging the wavelet coefficients of the ε-shifted copies of the input signal [191]. In their study, Coombes and his colleagues applied various levels of thresholds to the wavelet coefficients to find an optimal threshold that efficiently performs denoising but preserves the peaks in the mass spectra [170]. They reported that they identified more peaks of high quality from the denoised mass spectra than the raw mass spectra [170].

However, current noise reduction methods using variants of the WT attempt to simultaneously estimate and reduce noise on the basis of a single mass spectrum, which is an ill-posed problem [192]. This problem occurs when two unknown variables (noise and signal) are to be estimated from a single equation (the noisy signal). Thus, previous methods set criteria for quality assurance (*e.g.*, smoothness measurement) or made assumptions on the statistical properties of the signal or noise (*e.g.*, white Gaussian noise) [192]. Moreover, the non-white and non-stationary characteristics of chemical noise have not been considered in determining a universal threshold for all the decomposition levels. Thus, it is critical to investigate the characteristics of chemical noise in the wavelet domain and to design a denoising method based on this prior knowledge.

In this chapter, I propose a new thresholding method, adaptive thresholding using multiple realizations (ATMR). I investigated the time-frequency patterns of chemical noise in the wavelet domain with multiple realizations of a mixture of matrix (sinapinic acid) and several reference proteins. Using the patterns of chemical noise in the wavelet domain, I obtain the adaptive thresholds for each of the decomposition levels, which takes into account the non-white and non-stationary nature of chemical noise. The

threshold is extensively tested on another set of MALDI TOF mass spectra of different reference proteins to evaluate the denoising results. In addition to high-frequency noise reduction, I present a novel baseline correction algorithm, called the wavelet baseline correction (WBC) algorithm, which can eliminate the monotonically decreasing baseline in the highest level approximation of wavelet decomposition

4.2 BACKGROUND

In this section, I review fundamental theories for analysis of chemical noise using the WT. First, I introduce multiresolution analysis in signal processing. Multiresolution analysis enables decomposition of a complex function into multiple simpler forms using several resolutions so that they can be analyzed separately. The WT makes it possible to do multiresolution analysis by projecting a given signal onto multiple basis functions with frequency increasing by the power of 2. I give a brief mathematical review on the WT. In addition, I will introduce the SWT in more depth because my chemical noise analysis is based on the SWT. Finally, I review threshold estimation methods and thresholding methods in the wavelet domain that have been commonly performed in wavelet denoising.

4.2.1 Multiresolutional analysis and the wavelet transform

Unlike noise from instrumentation, chemical noise is non-stationary. In a stationary random process (*e.g.,* noise from instrumentation), at least its first order statistics (*e.g.,* mean) and second order statistics (*e.g.*, variance) are time-invariant. Based on this important property, autocorrelation and power spectral density can be obtained, which provides useful information for filter design. However, the statistical

characteristics of chemical noise are not stationary since matrix clusters, which are the primary source of the noise, are more likely to be formed in the low mass region than in the high mass region. For example, the mean of chemical noise over time is monotonically decreasing, which is time varying (Figure 4.1). Ordinary Fourier transform-based approaches are not suitable for non-stationary signal analysis because they can reveal only the power distribution over frequencies, but cannot describe how the power distribution changes with respect to time. Therefore, a more advanced signal processing technique than Fourier transform based methods for stationary random processes must be applied for spectral analysis of chemical noise.

Traditionally, non-stationary signals have been modeled and analyzed using the short-time Fourier transform (STFT). The STFT can show how frequency components vary over time by taking a Fourier transform of a windowed segment of a given signal and by moving the window over the time axis, resulting in a two dimensional representation of the signal (time vs. frequency). The STFT is mathematically represented as follows.

$$STFT(t', f) = \int_t \left[x(t) \cdot W(t - t') \right] \cdot e^{-j2\pi ft} dt \tag{4.1}$$

In Eq. 4.1, a moving window $W(t - t')$ is multiplied by the input signal $x(t)$ and a Fourier transform is performed at each translation t' of the window function. As a result, the frequency representation of $x(t)$ at t' is obtained. The energy distribution of the non-stationary signal over the time and frequency is called a "spectrogram," which corresponds to a power spectral density in spectral analysis of a stationary random signal.

While it is true that the STFT is very useful in analyzing non-stationary signals, deciding an optimal window size still remains an open problem of deciding an optimal window size. This problem is often referred to as the uncertainty principle of time and frequency [193], which says that there must exist a trade-off between the frequency resolution and the time resolution. Specifically, if the window size becomes larger, then the frequency resolution increases, but the time resolution decreases, and vice versa.

Multiresolution analysis (MRA) can alleviate this limit of the STFT without violating the uncertainty principle. Intuitively, for analyzing signal components with low frequencies, a larger window is used, and for analyzing signal components with high frequencies, a window with a smaller scale is employed. Thus, it enables analysis of a given signal at different frequency sub-bands with different resolutions. In particular, the WT performs MRA by dividing the time-frequency plane into dyadic tiles.

The wavelet transform enables a multi-resolution analysis. The wavelet transform provides good frequency resolution for low frequency signals and good time resolution for high frequency signals. This property of the wavelet transform is especially useful in analyzing high frequency components for short durations and low frequency components for long durations. In practical applications, the discrete wavelet transform is performed by using a filter bank. Using the filter bank, the given signal is decomposed into orthogonal sub-bands, each of which represents a specific frequency range over time. Mathematically, this signal decomposition using the discrete wavelet transform is described as

$$x(t) = \sum_{k,j \in Z} a_{j,k} \varphi\left(2^j t - k\right) \tag{4.2}$$

where φ denotes the basis (scaling function), k is the amount of shift in the time domain, and j is the level of decomposition. Just as the Fourier transform represents a signal as a set of coefficients indicating the amount of each basis function (sinusoid) it contains, the wavelet transform represents a signal as a set of coefficients ($a_{j,k}$) indicating how similar the signal is to the wavelet basis function at each scale. Under the assumption of orthogonality of a basis, the wavelet coefficients are derived by the function inner product as follows.

$$a_{j,k} = 2^{j/2} \int_{-\infty}^{\infty} x(t) \cdot \varphi^*\left(2^j t - k\right) dt \cong 2^{j/2} \sum_m x\left(\frac{m}{2}\right) \cdot \varphi^*\left(m - k\right) \tag{4.2}$$

From a multiresolution analysis point of view, a signal can be represented by the sub-spaces spanned by the orthogonal bases (scaling and wavelet functions). Suppose that a given function belongs to $L^2(R)$, where every function $f(t)$ has a well defined integral of the modulus of the function on the real line [194]. I define a sub-space V_0 of $L^2(R)$ as a space spanned by a set of scaling functions.

$$V_0 = \left\{ g(t) \mid g(t) = \sum_k a_k \varphi(t - k),\ \ k \in Z,\ \ \varphi \in L^2(R) \right\} \tag{4.3}$$

In Eq. 4.3, Z denotes the entire integer set. When I decrease the scale, I can increase the size of the subspace as follows.

$$V_j = \left\{ f(t) \mid f(t) = \sum_k a_k 2^{j/2} \varphi\left(2^j t - k\right),\ \ j,k \in Z,\ \ \varphi \in L^2(R) \right\} \tag{4.4}$$

Intuitively, when the time is multiplied by a power of 2, the resolution of the scaling function also increases by the power (if the power is less than 1, the resolution decreases). Thus, the sub-space spanned by the high-resolution scaling function can include more functions in $L^2(R)$; the size of the sub-space increases as well. Mathematically, this can be described as follows [194].

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset V_3 \cdots \tag{4.5}$$

However, instead of continually increasing the resolution of the scaling function to represent a given function $f(t)$, a set of functions that are orthogonal to the scaling functions are used to describe the complement of V_j. These functions are called as wavelet functions and are denoted by ψ. Suppose that the sub-space spanned by $2^{j/2}\psi(2^j - k)$ is W_j. Then, W_j is the orthogonal complement of V_j and it fills the difference between V_j and V_{j+1} [194].

$$V_{j+1} = V_j \oplus W_j \tag{4.6}$$

Eq. 4.6 can be recursively applied to derive a general representation of $L^2(R)$ based on the wavelet functions as in Eq. 4.7.

$$L^2(R) = V_0 \oplus W_1 \oplus W_2 \oplus W_3 \oplus W_4 \cdots \tag{4.7}$$

As a result, I obtain the wavelet representation of the given function as follows (Eq. 4.8 and 4.9).

$$f(t) = \sum_{k} c_{jo}(k) 2^{jo/2} \varphi(2^{jo} t - k) + \sum_{k} \sum_{j=j_o}^{\infty} d_j(k) 2^{j/2} \psi(2^j t - k) \tag{4.8}$$

$$\begin{aligned} c_j(k) &= \langle f(t), \varphi(t) \rangle = \int f(t) \cdot \varphi_{j,k}(t) dt \\ d_j(k) &= \langle f(t), \psi(t) \rangle = \int f(t) \cdot \psi_{j,k}(t) dt \end{aligned} \tag{4.9}$$

In Eq. 4.8, j_0 is usually set to 0, but it can be set to any arbitrary integer. $c_j(k)$ and $d_j(k)$ (Eq. 4.9) are referred to as approximation and detail coefficients, respectively. From Eq. 4.7 and 4.8, I know that the approximation coefficients $c_j(k)$ correspond to V_j and the detail $d_j(k)$ to W_j. Signal modification based on the DWT, such as denoising, is performed by manipulating the approximation or detail coefficients. The signal can be reconstructed from its wavelet coefficients. For example, in denoising using the DWT, reconstructing a measured signal after suitable modification of its wavelet coefficients can produce a version of the signal in which noise is suppressed and the true signal is enhanced. I refer the reader interested in the mathematical details of the WT to books and articles of this topic [190, 194, 195]. In general, the DWT is implemented using recursive low-pass and high-pass filtering followed by decimation (filter bank), and the time complexity of this filter-based algorithm is linear ($O(N)$) [194].

The discrete stationary wavelet transform (SWT) is often referred to as the undecimated discrete wavelet transform (UDWT). The DWT is a very powerful method for analyzing many types of stationary or non-stationary signals; however, since the DWT is not shift-invariant, a shift of the input signal in the time domain does not lead to a shift of the wavelet coefficients of the signal, resulting in inconsistent denoising performance for a shifted version of the input signal [191, 194, 196, 197]. This shift-

invariance problem can be overcome through the ε-decimated discrete wavelet transform.

Basically, the decimation step following the convolution of the DWT causes the DWT not to be shift-invariant [191, 197]. In a decimation, even or odd indexed elements of the input signal are chosen, resulting in potential bias in characterizing the signal in the wavelet domain. Pseudo-Gibbs phenomena exhibited around discontinuities after wavelet denoising are a good example of this [191]. In the ε-decimated DWT, all the possible choices of indices for decimation at every level of decomposition are taken into account. For a given maximum decomposition level J, ε is defined as a binary array as $\varepsilon = \left[\varepsilon^1\ \varepsilon^2\ \ldots \varepsilon^J\right]$, where $\varepsilon^j\ (1 \le j \le J)$ can have 0 or 1 and these binary numbers denote the choice of even or odd indices for decimation, respectively. Because the ε-decimated DWT performs the DWT considering every ε combination at every level of decomposition, the length of the wavelet coefficients does not decrease with decomposition. There are slightly different ways of performing the ε-decimated DWT. The simplest way is to directly run multiple DWTs; however, this direct method is extremely inefficient. The SWT can do this task in $O(n\log(n))$ time through a simple undecimated filtering [191, 197].

Conventional denoising strategies, thresholding, for the DWT are also applicable to denoising using the SWT. Actually, denoising using the SWT is known to be more robust and reliable than denoising using the ordinary DWT [198] since the final reconstructed signal from the wavelet coefficients is obtained by averaging the inverse transform of the ε-decimated (redundant) wavelet coefficients. Moreover, the shift invariance of the SWT makes it possible to estimate local thresholds for a given non-stationary noise and to effectively denoise it. This is referred to as interval dependent thresholding.

4.2.2 Denoising by thresholding

The main idea underlying wavelet-based de-noising methods is that the wavelet representation can separate the signal and the noise by distributing the energy of the noise over a large number of wavelet coefficients having small amplitudes and simultaneously distributing the energy of the signal over a small number of coefficients having larger amplitudes. Thus, the signal can be de-noised by applying a threshold to remove small wavelet coefficients in appropriate sub-bands.

The two most common types of thresholding operations are hard thresholding (Eq.4.10) and soft thresholding (Eq. 4.11) [193, 199].

$$y_{hard}(t) = \begin{cases} x(t), & |x(t)| > \lambda \\ 0, & |x(t)| < \lambda \end{cases} \tag{4.10}$$

$$y_{soft}(t) = \begin{cases} \mathrm{sgn}(x(t)) \cdot (|x(t)|), & |x(t)| > \lambda \\ 0, & |x(t)| < \lambda \end{cases} \tag{4.11}$$

In Eq. 4.10 and 4.11, $x(t)$ is the original wavelet coefficients of a given level of decomposition and λ denotes a threshold. $y_{hard}(t)$ and $y_{soft}(t)$ are the denoised wavelet coefficients by hard thresholding and soft thresholding, respectively. In general, it is known that hard thresholding keeps the edge information, but causes some artifacts while soft thresholding provides smoother denoising results than hard thresholding.

The selection of the threshold parameter λ is critical. If λ is too small as compared to the variance of noise, the residuals of denoising cause artifacts in the reconstructed signal. If λ is too large, important information in the signal is modified or

deleted. Many methods such as the universal threshold estimate, VisuShrink [188], and SureShrink [187] have been proposed to estimate the threshold parameter from an estimate of the noise variance.

VisuShrink is a thresholding method that has been most widely used for denoising [188]. The threshold in VisuShrink is defined as:

$$\lambda = \sigma\sqrt{2\log(N)} \tag{4.12}$$

where σ is an estimate of noise variance and N is the length of the signal. The threshold is selected based on the following mathematical results (Eq. 4.13) [200].

$$\lim_{N\to\infty} p\left(\max_{1\le i\le N}|z_i| < \sigma\sqrt{2\log N}\right) = 1 \tag{4.13}$$

z_i is a Gaussian stochastic process with zero mean and variance σ^2. The probability that the absolute value of z_i is smaller than $\sigma\sqrt{2\log(N)}$ converges to 1 as N goes to infinity [200]. In general, σ is not known; however, it can be estimated using the robust standard deviation estimate as in Eq. 4.14 [187]. .

$$\sigma = \frac{median_{k=1,2,\ldots,N/2}\left(\left|x_{1,k} - median_k\left(x_{1,k}\right)\right|\right)}{0.6745} \tag{4.14}$$

Since the median of absolute deviation is used to estimate the standard deviation (σ) of noise z_i, some outliers can be effectively rejected from the estimate of σ.

VisuShrink sometimes overestimates the threshold; it may remove signal components with low strength as well as noise. To alleviate this problem, Donoho and Johnstone designed SureShrink, which adaptively estimates a threshold level by minimizing the Stein unbiased estimate of risk (SURE) [187, 200]. This algorithm selects

the optimal threshold λ from $\left[0, \sigma\sqrt{2\log N}\right]$ that gives the minimum value of the mean-squared soft-thresholding error [187, 200].

Minimax thresholding estimates a threshold that minimizes the maximum risk of the $L^2(R)$ in estimating a function $f(t)$ [188, 200]. Unlike VisuShrink, there is no closed-form solution for estimating the minimax threshold and the threshold is numerically solved. However, the minimax threshold asymptotically converges to the VisuShrink threshold $\sigma\sqrt{2\log(N)}$ [188, 200].

Direct application of the thresholding methods described above for reducing chemical noise is not reasonable because the methods were designed under the assumption that the noise is additive, Gaussian, and white. However, chemical noise cannot be modeled as sich; hence, I attempted to estimate a threshold that can describe the statistical characteristics of chemical noise using multiple realizations.

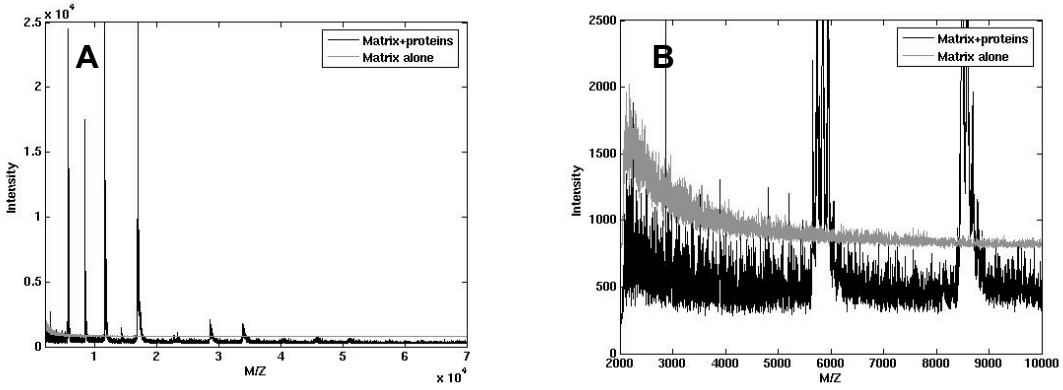

Figure 4.2: Mass spectra of matrix material alone and matrix plus several reference proteins (gray and black, respectively). B is a zoomed view of the low mass region of A. The matrix alone does not behave like the matrix plus proteins in MALDI TOF mass spectrometry.

Table 4.1: The sets of mass spectra used for the analysis of chemical noise. TR denotes training set and TS denotes test set. The thresholds obtained from TR sets were tested on their corresponding TS sets.

	TR 1	TS 1-1	TS 1-2	TR 2	TS 2
Reference Proteins	Insulin Thioredoxin Myoglobin	Ovalbumin Lysozyme C Insulin B chain	Ovalbumin Beta casein Ribonuclease B Cytochrome C	Insulin Thioredoxin Myoblobin	Ovalbumin Lysozyme C Insulin B chain Beta casein Ribonuclease B Cytochrome C
Set Size	20	30	20	10	42

4.3 MATERIALS AND METHODS

4.3.1 Data sets

A series of controlled experiments were performed in order to statistically characterize the noise from chemical impurities. However, the noise from chemical impurities cannot be completely isolated since spectra from any experiment exhibit noise due to instrumentation. However, since noise from instrumentation does not significantly affect the quality of a mass spectrum (Chapter 3), measurements of the chemical noise component are made in the presence of noise from instrumentation. Moreover, noise from random ion motions also exists because real materials (*i.e.*, matrix material and reference proteins) were used for the measurements. In fact, I observed that noise in the high mass region looks like noise from instrumentation; thus, my wavelet denoising also removes noise from instrumentation.

Since the largest component of the chemical noise arises from protonated clusters of matrix material, it might seem that the ideal experiment would be to collect mass spectra of matrix material alone. In practice, however, matrix material alone does not behave the same way as matrix material in the presence of protein (Figure 4.2). Figure 4.2 A shows an example mass spectrum of matrix material alone and B an example mass spectrum of matrix material plus several reference proteins. These two mass spectra were measured with the same instrumentation parameters. However, as shown in Figure 4.2, even by visual inspection, it is apparent that the noise properties (*e.g.*, noise variance and baseline) are different. Thus, I used only the mass spectra of matrix material plus several reference proteins to characterize chemical noise.

Table 4.1 summarizes the sets of MALDI TOF mass spectra that were used for the analysis of chemical noise. The data were acquired using a Voyager STR MALDI

TOF instrument (Applied Biosystems, Framingham, MA, USA), located at the Moffitt Cancer Center (MCC). The matrix consisted of a 28 mg/ml solution of sinapinic acid (Sigma-Aldrich, St. Louis, MO, USA) in 50% acetonitrile. The matrix was mixed 1:1 with the reference proteins in 1% acetonitrile and 0.1% formic acid. The choice of reference proteins was made based on three factors: (a) mass in the range of interest for protein profiling, (b) stable in solution, and (c) commercially available. In this study, insulin (5,080 Da), thioredoxin (12,000 Da), myoglobin (16,952 Da), ovalbumin (42,000 Da), lysozyme C (14,000 Da), insulin B chain (3,500 Da), beta casein (24,000 Da), ribonuclease B (14,700 Da), and cytochrome C (12,300 Da) were used as reference proteins (Table 4.1). The training sets (TR 1 and TR 2 in Table 5.1) were made from insulin, thioredoxin, and myoblobin, and the test sets (TS 1-1, TS 1-2 and TS 2) were created from combinations of the other proteins (Table 4.1). For TR 1, TS 1-1, and TS 1-2, the accelerating voltage was 25,000 V with 94% of grid voltage. The delay time was set to 400 ns and the bin size was 4 ns. For TR 2 and TS 2, some of the instrumentation parameters were slight changed by the operator several measurements to increase the qualitity of the mass spectra; 93.5% of the grid voltage and 600 ns of delay time were selected. It is common for the operator to tune the instrumentation parameters to have clear peak shapes. The laser intensity was varied between 2,300 and 2,600 to yield mass spectra of high quality. Although I could not find significant visual differences between these two groups of sets of mass spectra (*i.e.*, TR 1, TS 1-1, and TS 1-2 vs. TR2 and TS2), I performed separate experiments to prevent some unknown non-homogeneity between the two groups from interfering in my analysis.

4.3.2 Denoising using multiple mass spectra of chemical noise

The MATLAB® 7 Wavelet Toolbox was used to develop the adaptive thresholding using multiple realizations (ATMR) and wavelet baseline correction (WBC) algorithms. In the ATMR, the mass spectra in the training set (*e.g.*, TR 1 or TR 2) were decomposed by the SWT into approximations and details. In general, the choice of a suitable level of decomposition and wavelet basis depends on the signal and experience.

The level of decomposition was empirically determined to be eight (*i.e.*, $J = 8$) so that the highest level approximation could clearly demonstrate the monotonically decreasing baseline. As a wavelet basis, the Haar wavelet (Figure 4.3) was employed in this study because the use of this wavelet function clearly reveals the shape of the monotonically decreasing baseline as compared to the other wavelet functions we explored.

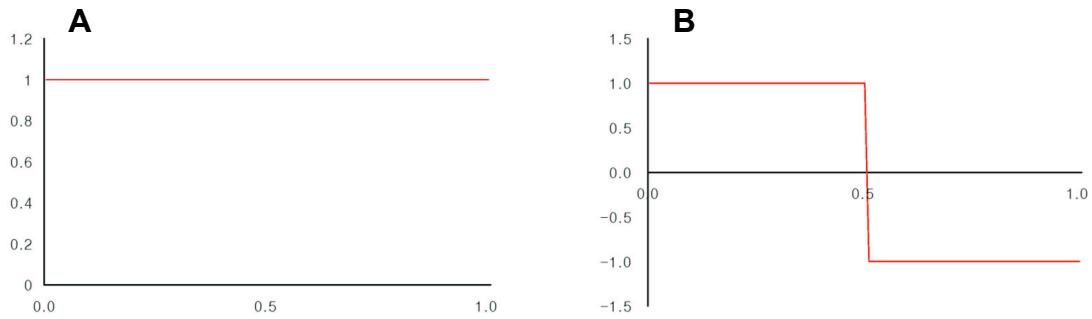

Figure 4.3: The Haar scaling (φ) and wavelet (ψ) functions (A and B respectively). These two functions are used to obtain the approximations and details of the wavelet decomposition respectively.

High frequency noise reduction and baseline elimination were performed using the details and approximations of the decomposed mass spectra respectively. In high frequency noise analysis, the noise variance of each detail is estimated for determining a threshold. Because the noise variance of chemical noise varies over time (*i.e.*, non-stationary), the details were evenly divided with an interval of 512 time points and then the noise variance of each interval was estimated using Eq. 4.14. The length of the interval was empirically selected such that the wavelet coefficients within the interval seem to be stationary. The threshold for each interval was calculated using Eq. 4.12, where the σ is the estimate of the noise variance of each interval and N is set to the length of the detail.

The median among the thresholds obtained from multiple realizations was selected as a robust threshold. For example, if there are 10 mass spectra of chemical noise in the training set, every interval will have 10 threshold estimates obtained from the 10

mass spectra. Then, the median in magnitude among those thresholds estimates is selected as the threshold for the interval.

As can be seen in Figure 4.4 A, the thresholds were over-estimated where the reference proteins were located. Moreover, the thresholds might reflect subtle changes in the noise variance. Therefore, a robust non-linear regression method was employed to reject the effect of the reference proteins and to smooth the thresholds. After testing a variety of regression functions such as linear and high-order polynomial functions, a two-term exponential function, $ae^{bx} + ce^{dx}$, was selected to model the pattern of the threshold of the exponential decrease in the low mass region and the plateaus in the high mass region (Figure 4.4 B). The use of a robust regression method makes it possible to reject the overestimated thresholds due to the reference proteins. The main disadvantage of least square regression is that the regression result is highly affected by outliers because outliers produce large squared errors and these large squared errors make the outliers more influential in the regression process than is appropriate. Robust regression with bisquare weights alleviates the influence of outliers by iteratively assigning small weights to the outliers during the regression process [201]. This algorithm calculates the weights for the outliers as the ratios of their residuals and the median absolute deviation of the residuals. The sample points with ratios larger than 1 are considered to be outliers and are ignored [201].

Finally, the threshold was zero-order interpolated over the entire region. The mass spectra in the test set (*e.g.,* TS 1-1, TS 1-2, or TS 2) were decomposed into the same number of levels and with the same wavelet. Then the decomposed details of the mass spectra were denoised with the threshold obtained from the training data set. In this study, soft thresholding was used because soft thresholding produces smoother curve shapes than hard thresholding.

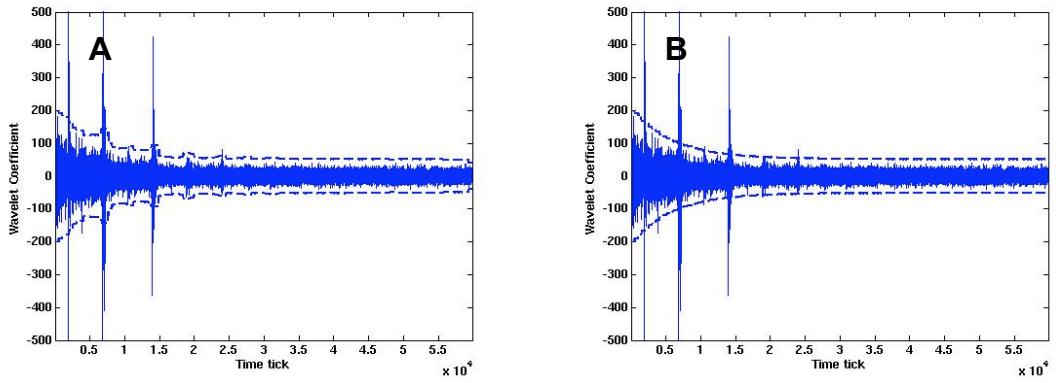

Figure 4.4: The 1 level detail of an example mass spectrum in TR 1 and its threshold estimates (dashed lines). The detail was divided into small intervals of 512 time ticks and a local threshold was estimated within each interval (A). Because of the reference proteins, the local thresholds were over-estimated in some regions. Robust non-linear regression analysis was performed using a two term exponential function, $ae^{bx} + ce^{dx}$, in order to obtain a smooth threshold (B).

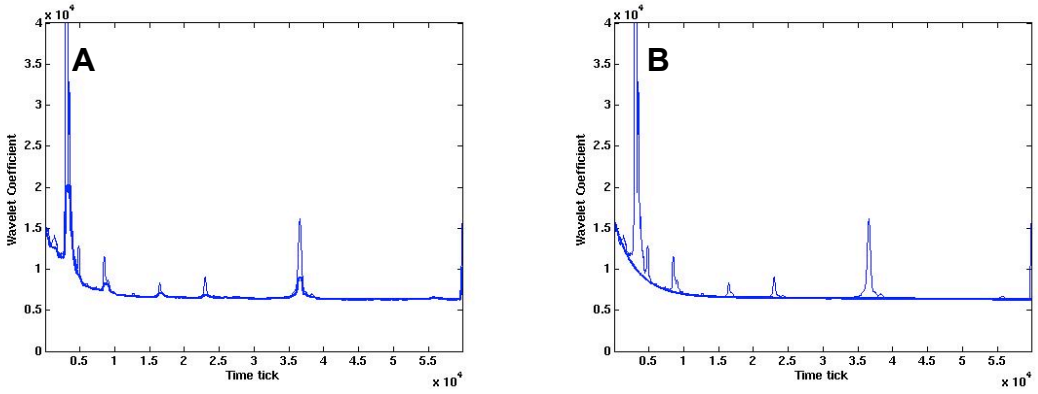

Figure 4.5: The 8 level approximation of an example mass spectrum in TS 1-1 and its baseline estimates (tick solid lines). The approximation was divided into small intervals of 100 time ticks and a crude baseline estimate was estimated with local minima of the intervals (A). Because of the reference proteins, the baseline was over-estimated in some regions. Robust non-linear regression analysis was performed using a two term exponential function, $ae^{bx} + ce^{dx}$, in order to reject the over-expressed values and smooth the baseline (B).

The baseline was eliminated at the highest level approximation (Figure 4.5) using the WBC algorithm. The baseline is clearly seen in the approximation without high-frequency components, which makes it easier to estimate the baseline in the wavelet domain than in the original signal. Unlike threshold estimation in high-frequency noise reduction, baseline estimation was performed on individual mass spectra since the baseline depends on the particular proteins contained in the sample.

In order to estimate the baseline, the entire mass (time) range was evenly divided into intervals, whose width was empirically set to 100. A more narrow width enables to obtain a finer baseline estimate. The lowest value of each interval was selected as a representative point and then the robust non-linear regression of a two-term exponential function was applied to the representative points in order to obtain a smooth baseline. This baseline estimate was also zero-order to the entire mass (time) range. After estimating the baseline, the highest level approximation was corrected with the baseline in a similar way to soft thresholding. That is, the points larger than the baseline estimate are shrunk by the height of the baseline and those smaller than the baseline estimate are reduced to zero.

The denoised details and the baseline-corrected approximation were reconstructed using the inverse discrete stationary wavelet transform (ISWT). The denoised mass spectra were evaluated in terms of visual inspection, the number of peaks, and the S/N of the detected peaks. In order to identify peaks from mass spectra, I employed the peak detection algorithm developed by Coombes *et al* [103].

The ATMR and WBC algorithms were extensively compared with conventional thresholding and baseline elimination methods such as VisuShrink. The baseline correction and denoising results of these two methods were qualitatively compared through visual inspection. In addition, the denoising results were quantitatively evaluated

in terms of the number of detected peaks and the S/Ns of the detected peaks. For this purpose, the simultaneous peak detection and baseline correction (SPDBC) algorithm developed by Coombes *et al*. [103] was used. The SPDBC algorithm operates in the time domain (m/z domain). This algorithm was also employed to eliminate the baseline of the mass spectra denoised using VisuShrink because VisuShrink only reduces the high frequency. The parameters of the SPDBC algorithm are determined by the resolution ($m/\Delta m$) of the MALDI TOF mass spectrometer. According to the product manual, the resolution of the Voyager STR MALDI TOF instrument (Applied Biosystems, Framingham, MA, USA) can be larger than 1,000 for myoglobin (16,952 Da) in the linear mode. In this study, I assumed that the resolution is about 200 in order to avoid selecting small variations due to noise.

4.4 EXPERIMENTAL RESULTS

4.4.1 Chemical noise characterization using the SWT

Figure 4.6 shows example mass spectra of the training (TR 1, blue) and test data (TS 1-2, red) respectively. As can be see in Figure 4.6, the noise variance of the two mass spectra in the low mass region look similar, but their baseline shape are different around the peaks of the reference proteins. This observation supports my approach to denoising chemical noise in which the high frequency noise is reduced based on a threshold estimated from the multiple realizations, but the baseline is individually corrected. Similar observations were also made about other training and test sets.

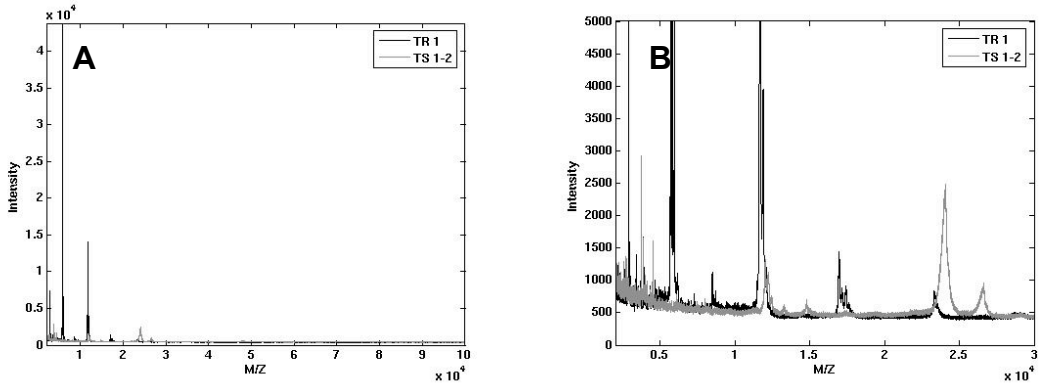

Figure 4.6: Example mass spectra of TR 1 and TS 1-2 (black and gray, respectively) (A) and their zoomed views of the low mass region (B). Overall, chemical noise is similarly expressed in TR 1 and TS 1-2 except for the peaks due to the reference proteins.

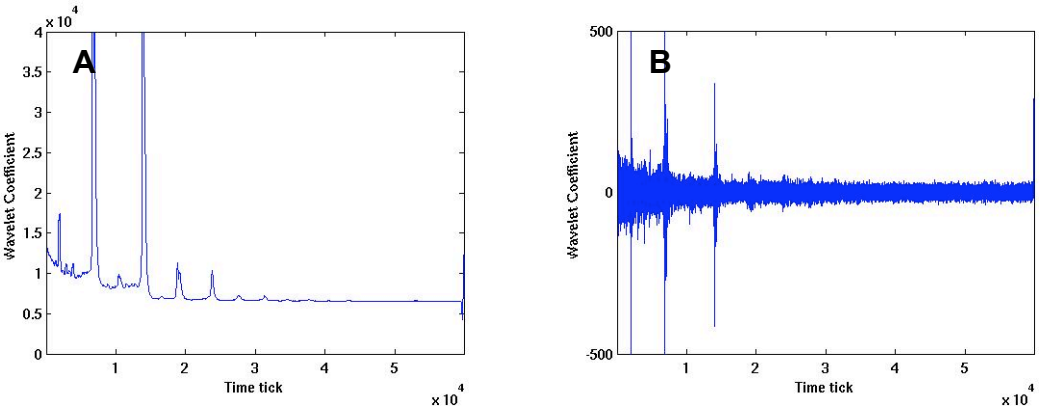

Figure 4.7: The 8 level approximation c_8 and 1 level detail d_1 of a mass spectra decomposed using the SWT (A and B respectively). c_8 shows the basic shapes of the peaks of the reference proteins and the monotonically decreasing baseline and d_1 displays the high frequency noise components.

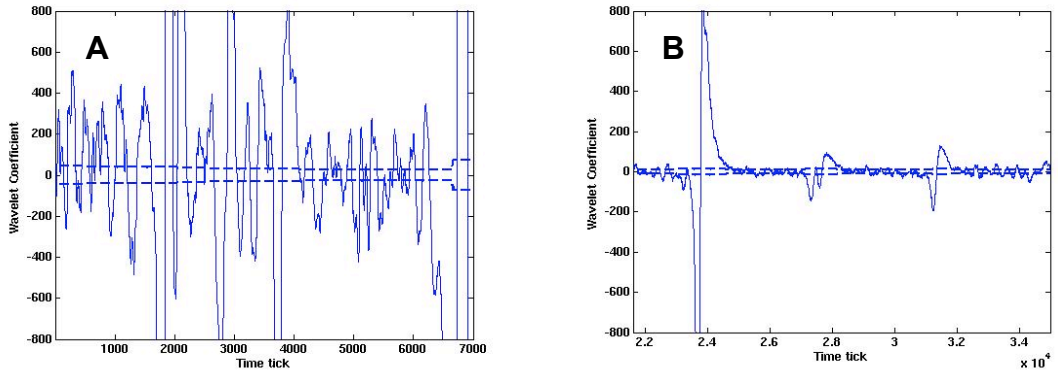

Figure 4.8: d_1 and d_8 of an example mass spectrum. The noise level of d_1 is marked as dashed lines. In the low mass region, the amplitude of d_8 is much larger than that of d_1 (A), while they do not look different in the high mass region (B). The sudden large variations were due to the reference proteins.

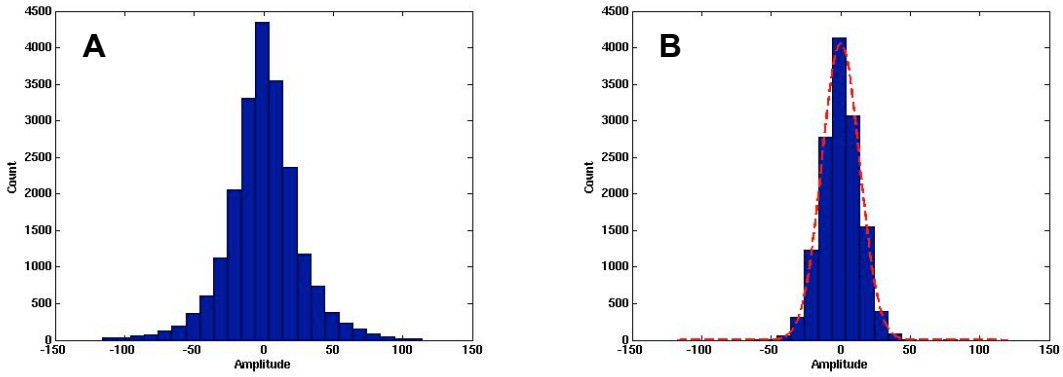

Figure 4.9: The histograms of the level 1 detail in the low mass region (m/z < 20,000) and high mass region (m/z > 20,000). The distribution of the high mass region could be modeled as a Gaussian distribution (dashed line in B). The distributions of low mass region had heavier tails on its both sides than a Gaussian distribution (A).

The mass spectra were decomposed to 8 levels (*i.e.*, $J = 8$) using the Haar wavelet. Figure 4.7 shows the 8 level approximation c_8 and the 1 level detail d_1 of an example mass spectrum (A and B respectively). As can be seen in Figure 4.7 A, the approximation shows the basic shapes of the peaks of the reference proteins and the monotonically decreasing baseline, which makes it possible to eliminate the baseline in the wavelet domain. The variance of the detail (Figure 4.7 B) also shows a monotonically decreasing shape over time. The large variations in the middle were caused by the reference proteins (Figure 4.7 B). Interestingly, this observation seems to support the noise model proposed in Chapter 3. According to the noise model (Eq. 3.1), the noise strength would increase with the heights of peaks because of the multiplicative noise component $R(t)$. Thus, the large variations over the chemical noise strength in the 1 level detail d_1 seem to account for the multiplicative noise in mass spectrometry. However, further studies need to be done to obtain the statistical characteristics of noise from random ion motion $R(t)$.

The 1 and 8 level details, d_1 and d_8, were studied to see if chemical noise could be modeled as white noise. Figure 4.8 compares d_1 and d_8 (dashed line and solid line) in a zoomed view. As can be seen in Figure 4.8, the amplitude of chemical noise in the low mass region (m/z < 20,000) increases along with the decomposition level. For instance, while the maximum deviation of d_1 is about 100, that of d_8 almost reaches up to 600 (Figure 4.9 A). In Figure 4.8, the *x* axis and *y* axis represent the time tick and the amplitude of the wavelet coefficients. The wavelet coefficients with large amplitudes in the middle of the x axis were formed by the reference proteins. It was observed that the amplitudes of the wavelet coefficients corresponding to the signals of the reference proteins and chemical noise grow together as the level increased. This observation shows that chemical noise is non-white and, in the wavelet domain, it behaves similarly to the

signals of the reference proteins. This is mainly because chemical noise originates from small organic molecules. In contrast, the noise level in the high mass region (m/z > 20,000) does not vary much across the levels (Figure 4.8 B), which shows that this noise can be modeled as white noise.

The distributions of chemical noise were also investigated. Figure 4.9 shows the distributions of the wavelet coefficients of the 1 level detail in the low mass and high mass regions (Figure 4.9 A and B respectively). As can be seen in Figure 4.9, the distribution of the high mass region could be modeled as a Gaussian distribution (the dashed line in Figure 4.9 B); however, the distribution of the low mass region does not appear to be Gaussian. This analysis confirms that noise in the high mass region is mainly due to instrumentation whereas chemical noise dominates in the low mass region.

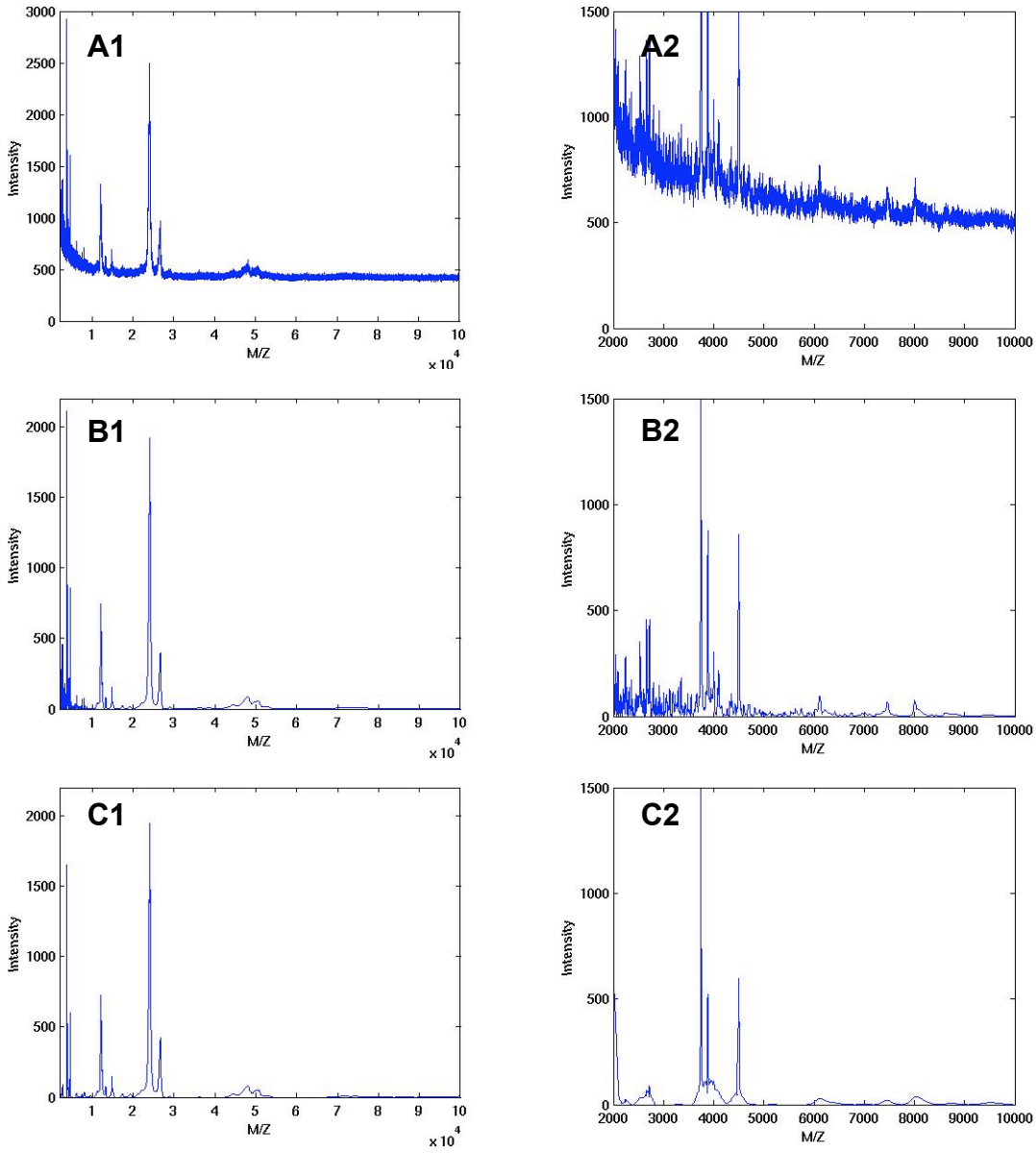

Figure 4.10: An example raw mass spectrum of TS 1-2 and the mass spectra denoised by VisuShrink and ATMR (A, B and C respectively). The images in the second column are the zoomed-in views of those in the first column.

Table 4.2: The number of detected peaks in the mass spectra in the test sets before and after denoising. The ATMR was compared with VisuShrink. The WBC algorithm was used with the ATMR for baseline correction. The SPDBC algorithm was employed to eliminate the baseline after denoising using VisuShrink.

	TS 1-1 (mean ± std)	TS 1-2 (mean ± std)	TS 2 (mean ± std)
Raw mass spectra	787±39	781±26	778±31
VisuShrink with SPDBC	287±30	289±22	284±25
ATMR with WBC	124±30	81±7	111±20

4.4.2 Results of baseline correction and denoising

The results of the ATMR and the VisuShrink algorithms were demonstrated in Figure 4.10. The image in the first row is an example raw mass spectrum of TS 1-2, and those in the second and third rows are the mass spectra denoised by ATMR and VisuShrink, respectively. As can be seen in Figure 4.10, the mass spectrum denoised by VisuShrink has artifacts due to the residuals from the denoising process (B2). By comparison, the mass spectrum denoised by ATMR shows a smoother curve than VisuShrink and has fewer artifacts in the peaks (C2). In particular, VisuShrink could not completely remove the chemical noise in the low mass region while my adaptive method successfully eliminated it (compare B2 and C2). However, both methods showed comparable denoising results in the high mass region. Similarly, it was observed that ATMR also outperformed VisuShrink in TS 1-1 and TS 2. Denoising by the VisuShrink method left large residuals in the low mass region as compared to ATMR, resulting in artifacts in the reconstructed mass spectra.

In addition, since the universal threshold tends to be overestimated relative to the true noise level in the high mass region, low peaks in the high mass region can be unnecessarily removed from the mass spectrum. However, use of the ATMR prevents this problem by adaptively estimating the threshold. The baseline line of each mass spectrum was also successfully eliminated by WBC in the 8 level approximation c_8. The SPDBC algorithm also showed comparable performance. No noticeable differences were identified between the mass spectra corrected by the SPDBC algorithm and my non-linear regression based method.

The efficacy of the ATMR algorithm was tested by investigating the quality of peaks detected by a peak detection algorithm. The mass spectrum is a mixture of the true peaks from the reference proteins and the false peaks from noise. Thus, if a denoising algorithm is successful, the number of peaks detected will be reduced in the denoised mass spectra compared to in the raw mass spectra because only a few reference proteins (*i.e.,* 4-7 proteins) were in the sample. Peak detection was done on the individual mass spectra of each data set and the number of detected peaks was counted. Table 4.2 shows the mean numbers of detected peaks from the mass spectra of each data set using VisuShrink and ATMR. The mean number of detected peaks from raw mass spectra of each data set is also given for comparison. As can be seen in Table 4.2, about 780 peaks were consistently detected from the mass spectra across the data sets. VisuShrink reduced the number of detected peaks to about 300; however, ATMR made it possible to detect far fewer peaks (about 80-120) than VisuShrink. This observation implies that ATMR effectively reduced the false positives in peak detection. In addition, the S/N can be used as a measure of the quality of detected peaks. The S/N is defined as the ratio between a peak height and the median absolute deviation of the mass spectrum near the peak, which is considered to be the noise level around the peak [103]. S/N was also dramatically

increased due to use of the ATMR algorithm. For example, the original S/N of the ovalbumin in TS 1-2 is only 8 without denoising; however, it increased to 10,000 when VisuShrink was used, and it is ever more improved to 30,000 by ATMR.

In summary, ATMR reduced chemical noise in mass spectrometry, which is dominant in the low mass region, while VisuShrink was only able to remove noise in the high mass region. The superiority of the ATMR algorithm over VisuShrink was confirmed by qualitative and quantitative evaluation of the denoising results. It was shown that the mass spectra denoised using my algorithm contained far fewer false positives than the raw mass spectra or the mass spectra denoised using VisuShrink. In addition, the S/Ns of the detected peaks were drastically improved by a factor of 4,000 when compared to VisuShrink.

4.5 DISCUSSIONS AND CONCLUSION

In MALDI TOF MS, noise hampers both humans and pattern recognition algorithms in observing the true biochemical nature of the sample. As an extension of the study of noise from instrumentation in Chapter 3, chemical noise was extensively analyzed using the WT in this chapter. The SWT was selected for the study because this variant of the WT can make the wavelet coefficients of a given signal shift-invariant, allowing for consistent denoising performance for similar signals shifted in time. This feature is important for denoising mass spectra in the sense that mass spectra could often be phase-shifted because of the innate mass error of the instrumentation. Moreover, the SWT enables interval dependent thresholding because the use of DWT decreases the time resolution of details and approximations by a factor of 2 as wavelet decomposition proceeds. This can be avoided by using the SWT.

Unlike noise from instrumentation, chemical noise originates from molecules with low molecular weights such as matrix material in MALDI TOF MS; thus, it cannot simply be modeled as Gaussian stationary white noise, which is a commonly employed assumption in denoising signals based on the WT. It was shown that chemical noise, particularly in the low mass region, is neither Gaussian nor white. Moreover, it is not stationary either since the noise level decreases over time (mass). Therefore, traditional denoising methods based on this assumption (*e.g.,* VisuShrink) are not appropriate to remove chemical noise. By contrast, ATMR can reflect the unique nature of chemical noise in estimating the optimal threshold based on the observation of multiple realizations of chemical noise. One should note that, in order to estimate the threshold for chemical noise that reflects the chemical noise characteristics under a certain set of instrumentation parameters, the ATMR algorithm needs mass spectra of matrix material plus several reference proteins under the same instrumentation parameters as those used for the sample parties to be denoised. I also developed a way of eliminating the monotonically decreasing baseline in MALDI TOF mass spectra using the WT. To the best of my knoweldge, there has been no study that attempted to estimate and remove the monotonically decreasing baseline in the wavelet domain. I successfully eliminated the baseline using non-linear regression with respect to local minima of the highest level approximation of wavelet decomposition.

The experimental results show that ATMR and WBC are superior to conventional denoising and baseline correction algorithms such as VisuShrink and SPDBC algorithm. In particular, the ATMR excellently removed noise in the low mass region while VisuShrink still left relatively large residuals. WBC showed comparable results to SPDBC by successfully eliminating the baseline of a mass spectrum; however, it would

be beneficial to use WBC because the high frequency noise reduction and baseline correction can be done together in the wavelet domain.

Non-linear regression based on a two term exponential function for eliminating the influence of the reference proteins and smoothing the threshold and baseline was very effective. However, the thresholds for the higher level (*e.g.*, 5-6) details tend to be overestimated in the low mass region, which might result in the loss of protein information. This phenomenon may be due to the fact that the signal amplitude of the reference proteins grows quickly in the higher level; the robust regression algorithm could not fully correct for this effect. This issue could particularly be resolved by using a different regression function instead of the two term exponential function. The two term exponential function is suitable for modeling the rapid decrease in the beginning or end of a signal. However, several huge outliers in the low mass region (*i.e.*, large amplitude of the reference proteins with low mass) can make the function estimate higher than it is supposed to be.

I also faced a similar issue in regard to the WBC algorithm. It was sometimes observed that the baseline obtained by WBC was overestimated in the middle mass region because of the high peaks of the reference proteins. As a result, small peaks in the middle mass region were eliminated by baseline correction. In particular, it is expected that the baseline is prone to be overestimated in the case that many high peaks are packed in a narrow region. This problem may be alleviated by sparsely sampling the points from that region and densely sampling from other regions to reduce the effect of the high peaks. However, automatically determining sampling intervals according to the peak height would be a technical challenge.

As mentioned in the previous section, the wavelet analysis also gave some clues to how to analyze noise from random ion motions. As can be seen in Figure 4.4 and 4.7,

the wavelet coefficients corresponding to the locations of the reference proteins had larger amplitudes than others. In particular, the large wavelet coefficients of the 1 level detail may originate the multiplicative noise (*i.e.,* noise from random ion motions) in mass spectrometry in the sense that the 1 level detail represents the noise frequency band. For denoising the multiplicative noise, the characterization of the noise must precede threshold estimation because the wavelet coefficients of the reference proteins have large amplitudes across all the decomposition levels and it should be investigated which levels are related to the multiplicative noise. Then, the noise should be carefully eliminated using a locally adaptive threshold, which adjusts the threshold depending on the change of the local noise level.

The conventional wavelet analysis recursively decomposes a given signal into an approximation and a detail. That is, the approximation keeps being split into the next-level approximation and detail until a proper level of decomposition is achieved, while the details are not decomposed further. Wavelet packet analysis decomposes the details as well as the approximations, providing more perspective on the signal. Thus, wavelet packet analysis is expected to more clearly reveal the characteristics of chemical noise or multiplicative noise in the wavelet domain than the traditional wavelet analysis.

In conclusion, the adaptive thresholding using multiple realizations of chemical noise and the wavelet baseline correction algorithms effectively removed chemical noise from mass spectra. In my study, ATMR and WBC yielded superior results to conventional wavelet denoising and baseline correction methods. Several suggestions were made for future extensions of my study to identify more realistic thresholds for higher level details and a baseline than the current method. In addition, the experimental results imply that multiplicative noise can also be removed through wavelet analysis, particularly wavelet packet analysis. It is also expected that wavelet packet

analysis would enables to precisely remove chemical noise by providing a more clear view of the time-frequency composition of chemical noise.

Chapter 5 GUILT-BY-ASSOCIATION FEATURE SELECTION: IDENTIFYING BIOMARKERS FROM PROTEOMIC PROFILES

5.1 INTRODUCTION

The goal of proteomic profiling by mass spectrometry is to identify proteins/peptides (biomarkers) that are expressed differently between different disease states, *e.g.*, control and disease, such that a diagnostic test can be developed. However, as discussed in Chapter 2, many bioinformatics challenges must be overcome in order to obtain reliable biomarkers for disease detection, prognosis, and treatment monitoring [102].

One critical issue is the so called *curse of dimensionality*, which refers to the performance degradation of classification due to having few data samples in a high-dimensional variable space [202]. From a machine learning perspective, the protein/peptide species in a protein profile are variables for classification, which are often referred to as features. In particular, the height or area of a peak is often taken as a feature. The number of samples required for maintaining the same sample density in a feature space increases exponentially when another dimension is added. In protein profiling using mass spectrometry, selecting effective features that separate diseased samples from healthy ones can directly lead to identifying potential biomarkers. This operation is referred to as *feature selection*. The development of feature selection algorithms that operate reliably and robustly under this condition is the most critical issue in biomarker identification.

Traditionally, feature selection algorithms are divided into three categories: filters, wrappers, or embedded methods [139]. No one feature selection method is superior to all the others; every method has its own advantages and disadvantages. For

example, filters usually run faster than wrappers, but wrappers are more likely to select features that can produce better classification results than filters. Ideally, a feature selection algorithm must be able to identify features that are not only discriminant but also independent features in order to achieve optimal performance [139, 203]. However, many feature selection algorithms, particularly filters, evaluate features solely in terms of individual discriminability. Although wrappers consider the interrelationships between features, in order for the process to be computationally tractable, a search algorithm such as forward selection or backward elimination must be used. The search algorithm iteratively forms the best subset of features that maximizes the performance of the embedded classifier. However, the search result could be biased because of the heuristic and greedy aspects of the search algorithm used in feature selection. For example, forward selection or backward elimination determines which feature is supposed to be selected or eliminated at the next stage based on the current set of features. Thus, it is difficult to guarantee the mutual independency of the selected features. Genetic algorithms attempts to avoid such bias by optimizing the selection of subset on the basis of natural selection and randomization [204, 205]; however, in general, the running time of a genetic algorithm is extremely long compared to other search algorithms.

In general, highly correlated features do not improve classification performance [139]. In a simulation study, Guyon and Elisseeff observed that adding a highly correlated feature did not contribute to the information gain for classification unless the direction of correlation was orthogonal to that of class separation [139]. In recognition of this problem, several machine learning studies have presented feature selection algorithms for identifying discriminant and independent features based on information theory.

Koller and Sahami defined discriminant and independent features as *relevant* features and attempted to obtain relevant features by eliminating features mutually independent of class labels (*irrelevant* features) and features interrelated with other features (*redundant* features) [203]. The principle of this algorithm is that irrelevant and redundant features are classified on the basis of expected mutual information with the class labels and Markov blanket estimates respectively, and then they are eliminated by backward elimination. Hall also developed a similar feature selection algorithm that evaluates the relevancy of features on the basis of a combined correlation measure, and searches for an optimal subset of features using a conventional search algorithm such as forward selection or backward elimination [206]. However, these feature selection algorithms involve a search process and repetitive measurement of feature relevancy, resulting in a significantly high computational burden. Moreover, relevancy measures based on information theory may not be suitable for features from proteomic profiling since these measures are based on the estimation of the joint probability density function (PDF), which for proteomic data is high dimensional and may be poorly defined given the typically moderate number of samples.

On the other hand, several other studies have attempted to choose relevant features by clustering redundant features. Mitra and Pal proposed a feature selection algorithm that clusters features together based on the k nearest neighbor principle, and picks the feature with the smallest distance from other features within the cluster as a representative [207]. Since this algorithm does not require a search process, its running time is expected to be shorter than those algorithms introduced above. In addition, the pairwise distance definition based on a correlation measure (correlation coefficient or covariance) between features is more robust than the independency measure based on joint PDF estimation in the analysis of proteomic profiling. However, in this algorithm,

the user must specify several parameter values (*e.g.,* the number of neighbors, k) for optimal operation. Moreover, merely choosing the least correlated features may decrease classification performance since discriminability is not taken into consideration.

In another study, an attempt was made to group correlated features using hierarchical clustering [208]. King used the correlation coefficients between features as a similarity measure, and clustered features based on their correlation coefficients in a step-wise manner. However, specific stopping criteria are not suggested; the algorithm simply continues the merging process until only two clusters remains. Obviously, this algorithm is not appropriate for analyzing complex data like protein profiles because even uncorrelated features can be forced to fall in the same cluster.

In this chapter, I present a method, referred to as guilt-by-association (GBA), which reveals the dependency structure of features in a dataset using a correlation measure and a clustering algorithm. In my present study, I employed the correlation coefficient and the agglomerative hierarchical clustering algorithm; however, other dependency measures or clustering algorithms could be used. It should be also noted that, unlike prior studies reviewed above, GBA operates in cooperation with a filter feature selection method, such as the two sample *t* test, in order to select features that are discriminant as well as independent. In a previous study, I tested the GBA on simulated proteomic data sets, and observed that the GBA successfully distinguished the relevant features from other redundant and irrelevant features [109]. In this chapter, I evaluated the efficacy of the GBA algorithm with a real SELDI TOF breast cancer data set that had been analyzed by Pusztai *et al.* [111]

5.2 MATERIALS AND METHODS

5.2.1 GBA Algorithm

MATLAB® 7 (The MathWorks, Inc., Natick, MA) was used throughout the study. Selecting an appropriate distance measure is an important issue for cluster analysis [207, 209]. In this study, I define the pairwise distance between two features, x_1 and x_2, as in Eq. 5.1 and Eq. 5.2.

$$1 - \left|\rho(x_1, x_2)\right| \tag{5.1}$$

$$\rho(x_1, x_2) = \frac{cov(x_1, x_2)}{\sqrt{var(x_1)var(x_2)}} \tag{5.2}$$

where *cov* and *var* represent the covariance of two features and variance of a feature respectively. This distance measure is not a metric because it cannot satisfy the triangular inequality. I took the absolute correlation coefficients in defining the distances since both positive and negative correlations indicate dependency between the variables. After obtaining the pairwise distances between the features, the GBA algorithm iteratively groups the closest ones using agglomerative hierarchical clustering. Single, complete, and average linkage were empirically compared, and the average linkage was selected because it showed robust performance in a simple simulation study.

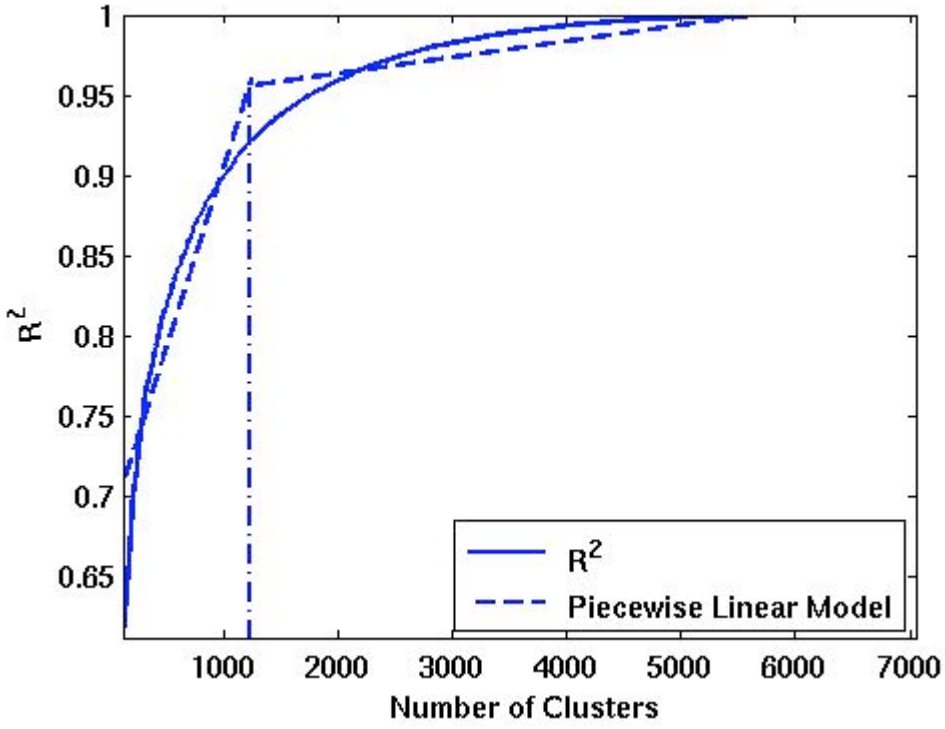

Figure 5.1: The R^2 measure of the breast SELDI Data (solid line). There are 7,052 features obtained directly from the baseline-eliminated and normalized spectra. The curve smoothly increases from 0 to 1 with the number of clusters. The optimal number of clusters is determined at the point where two piecewise linear regression lines (dashed lines) meet. These lines approximate the R^2 curve with the minimum-squared error. As a result, clustering stops at 1,111 clusters (dashed dot line).

As a clustering stopping criterion, R-squared (R^2) was employed. R^2 is defined as the ratio between the square sum of between clusters (SS_b) and the total square sum (SS_t) [210].

$$R^2 = \frac{SS_b}{SS_t} \tag{5.3}$$

R^2 *has* values from 0 (a single cluster of all the elements) to 1 (individual elements as clusters), and it monotonically increases in a concave form (Figure 5.1). Because R^2 measures the homogeneity of the clusters at every merging step, a sharp decrease in R^2 indicates that the merging process should be stopped. Figure 5.1 shows an example of R^2 (solid line) as a function of the number of clusters for the breast SELDI data set used in this study. As can be seen in Figure 5.1, the R^2 is relatively flat until the number of clusters reaches about 1,100 (vertical dashed dot line), and then decreases quickly. This implies that the smallest number of clusters that keeps the clusters homogenous is about 1,100. I identified this optimal stopping point using piecewise linear regression since simple differentiation of the curve with respect to the number of clusters did not clearly reveal the stopping point. As can be seen in Figure 5.1, the dashed lines are the two piecewise regression lines that give the minimum squared error of regression, and the point connecting these two lines was selected as the stopping point of clustering.

The features within a cluster are more dependent on each other than those that belong to different clusters; therefore, selecting representative features of the clusters can minimize the risk of choosing redundant features. In the GBA algorithm, representative selection is based on discrimination power, as measured using a metric such as the two sample *t* test. Moreover, as discussed in the previous section, GBA is designed as a preprocessing step to reduce the number of features based on feature redundancy prior to application of a feature selection routine that emphasizes discrimination power. Thus, the choice of the metric for selecting the representative feature of a cluster should be matched to the metric optimized in the subsequent feature selection routine. Figure 5.2 summarizes the GBA algorithm for feature selection.

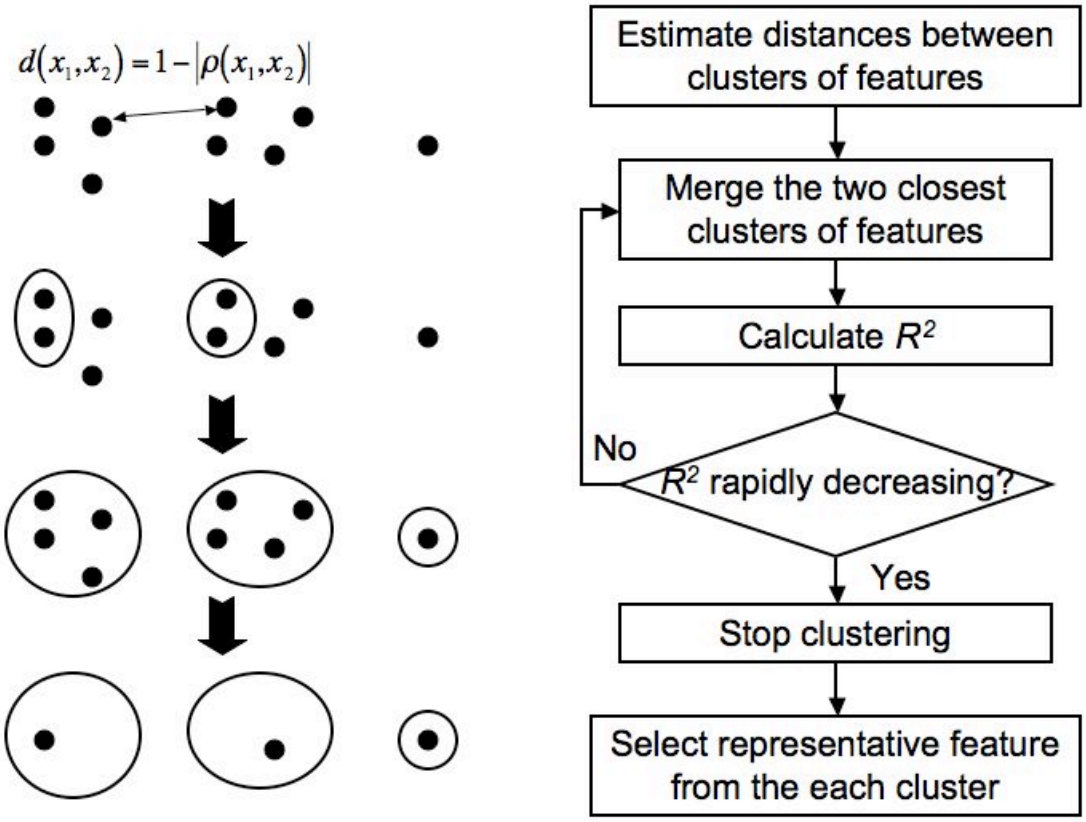

Figure 5.2: The GBA algorithm. The GBA consists of two parts: feature clustering and selecting representative features. The GBA calculates the pairwise distance between each pair of features, which is defined as $1-|\rho(x_1,x_2)|$, where $\rho(x_1,x_2)$ is the correlation estimate of x_1 and x_2. The agglomerative hierarchical clustering clusters features until the stopping criterion is satisfied. A standard filter method (*e.g.*, two sample *t* test or individual ROC analysis) serves selecting representative features from the formed clusters.

Table 5.1: The four data sets provided by Department of Biostatistics and Applied Mathematic at The University of Texas M. D. Anderson Cancer Center. In this study, 'March 03 Low Mass Scans' was used to test the efficacy of GBA.

Data Sets	Samples/Spectra			
	Pre-chemotherapy		Post-chemotherapy	
	Cancer	Normal	Cancer	Normal
March 03 Low Mass	25/47	15/29	26/49	15/28
March 03 High Mass	26/47	15/29	26/49	15/28
June 03 Low Mass	42/84	0/0	30/60	0/0
June 03 High Mass	42/84	0/0	30/60	0/0

5.2.2 Data set

SELDI TOF spectra of plasma from women with and without breast cancer were used to test the efficacy of GBA. This data set is described by Pusztai [111] and additional information is available on the website of the Department of Biostatistics and Applied Mathematics at The University of Texas M. D. Anderson Cancer Center (http://bioinformatics.mdanderson.org/pubdata.html). The mass spectra themselves can also be downloaded from the same location. I briefly summarize the pertinent aspects of the data. An IMAC-Cu metal binding chip was used to produce the mass spectra. There are a total of four data sets on the website, which were measured at two different time points (March 03 and June 03) and optimized to two different mass scanning regions (low and high mass regions) (Table 5.1). For my study, I selected 'March03 Low Mass Scans', which consists of mass spectra of plasma from 15 normal controls and 26 breast cancer patients over a low mass range (10-96,900 Da). For quality assurance, all of the samples were run in duplicate. This study used the normal mass spectra and pre-chemotherapy mass spectra for the cancer patients. Three mass spectra of the healthy controls and five

mass spectra of the cancer patients were excluded from the data set by the producers because the spectra showed poor signal quality during instrumentation. In addition, I also exclude from my analysis three mass spectra that do not have their pairs marked as high-quality signal in the 'March03 High Mass Scan' set. Also, one cancer sample was excluded because it did not have a pre-chemotherapy measurement. Thus, I used 29 mass spectra of 15 normal controls and 47 mass spectra of 25 cancer patients in my analysis (Table 5.1). Since the biochemical processes of sample preparation and detailed clinical information of the samples are not within the scope of this chapter, I refer the reader interested in these topics to Pusztai *et al*. [111] and the information files posted on their website.

5.2.3 Data processing

Data preprocessing consists of three steps: baseline elimination and peak detection, normalization, and peak alignment across the spectra. The baseline of each mass spectrum was eliminated using the simultaneous peak detection and baseline correction (SPDBC) algorithm developed by Coombes *et al* [103]. This algorithm detects the peaks in a given mass spectrum within a moving window, estimates the baseline using the peak-free leftover spectra in the window, and subtracts it from the original mass spectrum. During peak detection, peaks within mass resolution ($m/\Delta m$) of 200, which is set smaller than the typical mass resolution of SELDI TOF (300-400) to raise peak detection sensitivity, are combined into a one peak. In the SPDBC algorithm the signal to noise ratio (S/N) is defined as the ratio between the height of a peak and the median absolute deviation (MAD) from the median in the window around the peak [103]. The peaks with S/Ns less than 3 are eliminated from the peak list.

After baseline elimination and peak detection, the mass spectra were normalized with respect to the total ion current (TIC) of the mass region of 2,000-12,000 Da since the low mass region (< 2,000 Da) of SELDI TOF is generally corrupted by chemical noise [42, 51], and the high mass region (> 12,000 Da) has very low intensity peaks. Thus, only peaks within the middle mass region (2,000-12,000 Da) were considered for further analysis.

Since peaks representing a certain protein do not occur at exactly the same mass location across multiple mass spectra, peaks must be grouped together based on the mass accuracy of the mass spectra to make inferences about overall patterns across multiple spectra. This process is referred to as peak alignment. The mass accuracy of this breast SELDI TOF data set does not appear to have been reported. I assumed that the mass accuracy was at least 0.1%, a value reported frequently in other studies [211, 212]. The peak alignment process yielded 107 peaks.

Most of the samples have more than one mass spectrum because the samples were run in a duplicate manner. Thus, as Pusztai *et al*. [111] did in their analysis, I generated a consensus peak list by averaging each pair of replicate peak lists of the same sample.

Peak detection and peak alignment reduce the number of features by selecting the representative peaks of protein species from raw mass spectra. This process is similar to GBA in that the peaks within the resolution or mass accuracy would be highly correlated and peak detection and peak alignment removes those redundancies. However, while peak detection and peak alignment perform this function only on features of similar m/z values, GBA can provide a global view of the relationships between all the features. Thus, in addition to applying GBA after peak detection and peak alignment, I also investigated the direct application of GBA to the middle mass region (2,000-12,000 Da) of normalized mass spectra, which consists of 7,052 m/z points (features).

5.2.4 Experimental design

In this study, the two sample *t* test was used as a simple filter method for feature selection. As in Pusztai *et al.*'s analysis, a constant (0.5) was added to the denominator of the t statistic so that large *t* values are not be generated simply because the variances of the two groups (cancer and normal) are small due to very low intensities near zero, rather than features' true discrimination ability [111] (Eq. 5.4).

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\mathrm{var}(\bar{x}) + \mathrm{var}(\bar{y}) + k}} \tag{5.4}$$

where $\bar{x}$ and $\bar{y}$ are the mean estimates of two groups and $\mathrm{var}(\bar{x})$ and $\mathrm{var}(\bar{y})$ are the variance estimates of $\bar{x}$ and $\bar{y}$. k is a constant, 0.5 in this study, and this value was also used by Pusztai *et al.* [111]. I did not try to optimize this constant for my normalized mass spectra.

Two feature selection methods, *t* test alone and *t* test combined with GBA, were compared with different numbers of features selected. In the case of *t* test alone, features were simply selected according to their discrimination abilities; features with large absolute *t* values were chosen. However, in the case of GBA combined with *t* test, a feature with a large absolute *t* value was not selected if it was grouped with another feature with larger absolute *t* value. I evaluated the selected features in terms of the area under the Receiver Operating Characteristic curve (AUC) of a logistic regression classifier trained with those features.

The AUC of a classifier was obtained via 10-fold cross-validation to alleviate the difficulty of accurate evaluation due to the small data set size (40 samples). The entire

data set was randomly split into 10 non-overlapping partitions. Features were selected and a classifier was trained on nine partitions, and the classifier output values of the samples belonging to the remaining partition were obtained using the trained classifier. This process was repeated such that every partition was withheld once, and then a single AUC was estimated based on the classifier output values from the 10-fold cross-validation.

The whole process was repeated with different numbers of features selected (1-20) to see the performance change of GBA over the number of features. The range was set based on the numbers of potential biomarkers reported in similar studies [37, 40, 58, 80, 83, 86, 137]. As mentioned in the data preprocessing section, the same experiments were done both with peaks identified by preprocessing (107 features) and directly on the normalized mass spectra (7,052 features). Figure 5.3 summarizes the overall experimental process.

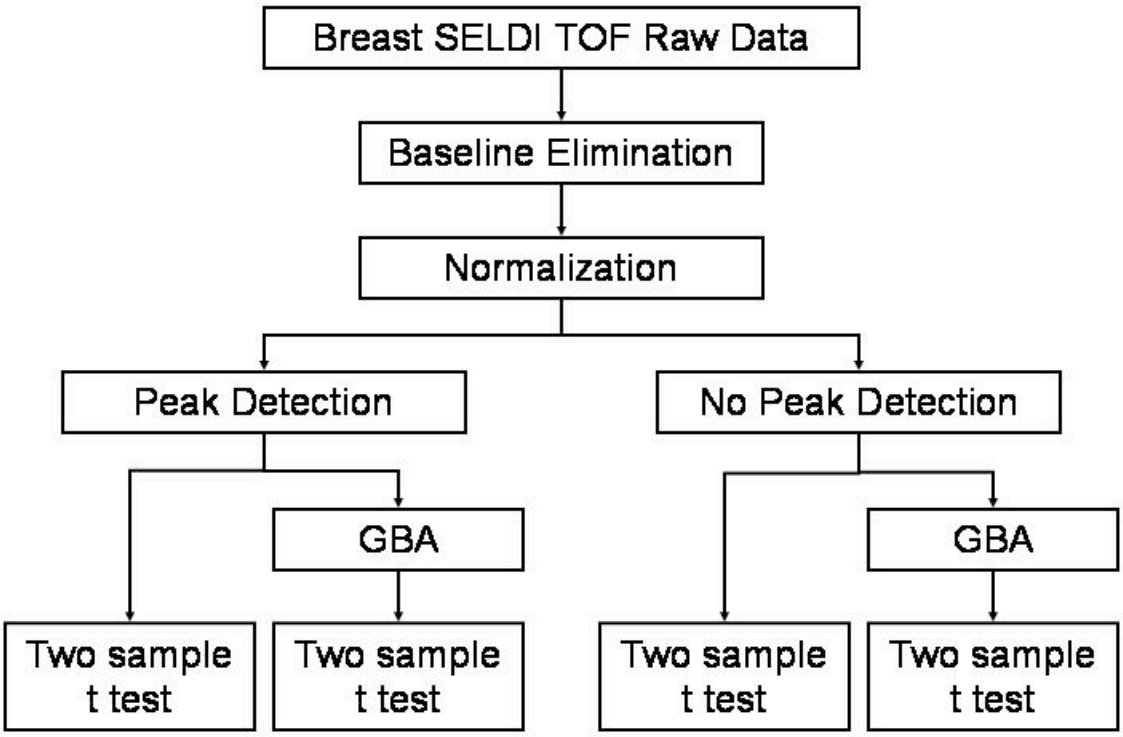

Figure 5.3: The summary of the experimental design. I tested the efficacy of the GBA with the peaks detected by the SPDBC peak detection algorithm [103], and the mass spectra without peak detection. Features were selected by *t* test alone and *t* test with GBA respectively and the selected features were evaluated in terms of the AUC of a logistic regression classifier trained with those features. The AUC of a classifier was obtained via 10-fold cross-validation to alleviate the difficulty of accurate evaluation due to the small data set size (40 samples).

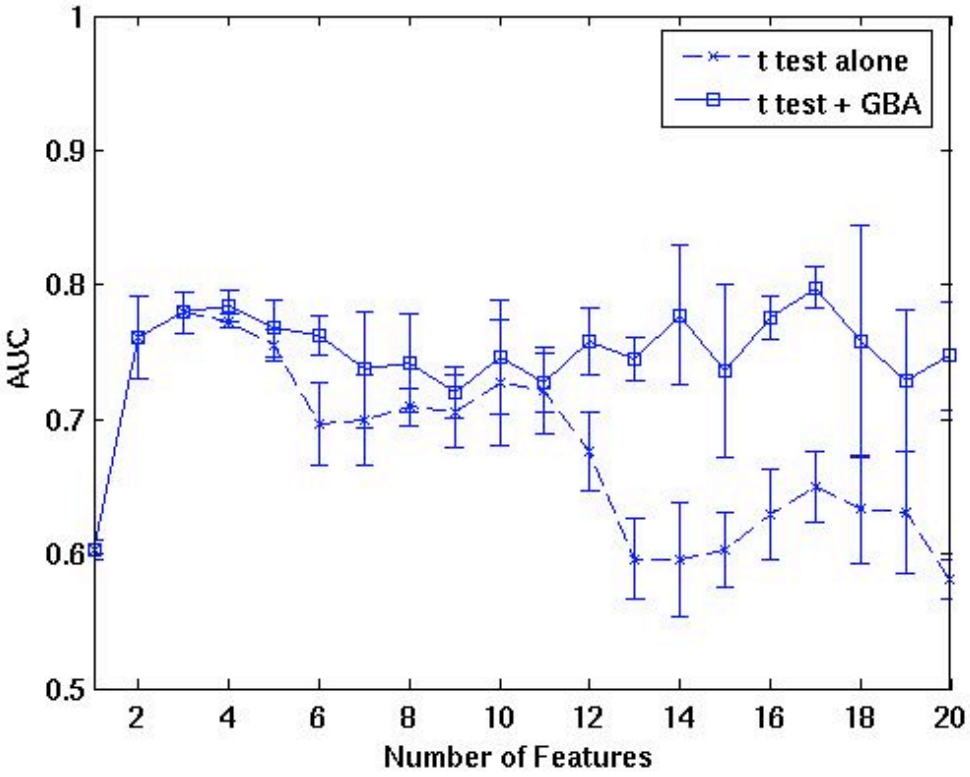

Figure 5.4: The AUCs of the logistic classifiers trained on the features selected by two sample *t* test alone and those by *t* test with GBA with the peaks identified by preprocessing. The error bars represent the standard deviations obtained from three different 10-fold cross-validation runs. The comparison of *t* test alone and *t* test with GBA was repeated with the number of selected features varying from 1 to 20 (horizontal axis). *t* test with GBA shows better than or comparable to *t* test alone. The curve of GBA more quickly arrives at a higher value than *t* test alone.

5.3 EXPERIMENTAL RESULTS

5.3.1 Two sample *t* test vs. *t* test with GBA on the peaks identified by preprocessing

The discrimination abilities of the individual features were measured using Eq. 5.4. In the case of *t* test alone, the features were sorted by their absolute *t* values, and the top *N* features were used for building a classifier. Since this feature selection process was performed inside of the cross-validation loop, the *t* values of the features were separately estimated for every run of the cross-validation. Similarly, in the case of *t* test with GBA, the representative features of the feature clusters identified by GBA were evaluated and representatives with the N largest *t* values were selected for classification.

Figure 5.4 shows the AUCs of *t* test alone and of *t* test with GBA over various numbers of features obtained from the peaks identified by preprocessing. The *x* axis represents the number of selected features, and the y axis represents the average AUCs (solid and dashed lines) and the standard deviations (error bars) obtained from three different cross-validations. It was observed that *t* test alone and *t* test with GBA produced equivalent performance (0.60-0.78) until about three features were selected, and then *t* test with GBA began to outperform *t* test alone. Considering the standard deviations, the difference between these two AUC curves of *t* test alone and *t* test with GBA is moderate when a small number of features are used, but *t* test with GBA is clearly superior as more features are included. On average, the AUCs of GBA reached the maximum ($\approx$ 0.79) when four features were included, and then remained above 0.72 whereas the AUCs of *t* test alone decreased to about 0.67 when the 6^{th}-9^{th} features were included, and then sharply fell again to about 0.6 when the 13^{th}-16^{th} features were added. This observation implies that the 6^{th}-9^{th} and 13^{th}-16^{th} features selected by *t* test alone are highly correlated with those previously selected, resulting in the reduction of classification performance. In

contrast, the AUC of t test with GBA is stable, remaining over 0.72, which suggests that the correlations among the features selected by *t* test with GBA were kept low by GBA.

In order to confirm my observation, I investigated how the features selected by *t* test alone were associated with each other and how well GBA identified the association structure of the features. Table 5.2 lists the top 20 features as ranked by *t* test alone over 10-fold cross-validation. The first column of Table 5.2 represents the mass to charge ratios (m/z) of the selected features, and the second column shows their mean *t* values estimated from the 10-fold cross-validation. Each pair of numbers from the 3rd column of Table 5.2 indicates which position a feature was ranked at by *t* test alone (outside of the parentheses), and how many times the feature was ranked at the position (inside of the parentheses) during the 10-fold cross-validation. For example, m/z 7766.9 in the first row of Table 5.2 has a mean *t* value of -0.00313 and was ranked #1 by *t* test alone six times (*i.e.*, 1(6)) and #2 four times (*i.e.*, 2(4)) during the 10-fold cross-validation. The *t* values in the last column of Table 5.3 are fairly low (< 0.003) compared to the *t* values reported by Pusztai *et al.* (< 4) [111] This may be because a large value (0.5) was selected for *k* of Eq. 5.4 as compared to the variances of peaks. I normalized the mass spectra with respect to the TIC of the mass region of 1,200-12,000 Da while Pusztai *et al.* selected a different mass region for normalization [111]. Therefore, the variances of peaks may have changed after normalization, resulting in low *t* values. However, the order of the features is not be affected by the choice of *k*.

Table 5.3 shows how these features were clustered by GBA. The features in each row of Table 5.3 belong to the same cluster, and the feature with an asterisk has the highest *t* value among those in the cluster (representative feature). The pairwise distances among the features belonging to a cluster are within from 0.02 to 0.23 (correlation of

0.98-0.77) while those among features in different clusters range from 0.25 to 0.93 (correlation of 0.75-0.07). As can be seen in Table 5.2, m/z 7766.9, 4443.8, and 2026.0 were ranked 1st-3rd, and these features were not significantly correlated together (Table 5.3). They belong to three different clusters (Clusters 1-3). GBA does not improve feature selection in the case that the top N features are independent and the N (or less than N) features already show the best classification.

However, when they can be grouped into a smaller number of clusters than N (*i.e.*, in the cases that some or all of the N features were significantly correlated), the combination of *t* test and GBA makes it possible to select more powerful feature sets than *t* test alone. In Figure 5.4, *t* test with GBA began to produce higher AUCs than *t* test alone when more than four features were selected; although the improvement is not large at first. m/z 4289.5 and 8939.6 were ranked by *t* test at the 4th or 5th positions (mean *t* value of 0.00145 and 0.00116 respectively in Table 5.2). While m/z 4289.5 was not redundant to the top three features (m/z 7766.9, 4443.8, and 2026.0), m/z 8939.6 was highly correlated with m/z 4443.8 (Cluster 2 in Table 5.3). Thus, including m/z 8939.6 decreased the AUC. In contrast, when GBA was used with *t* test, m/z 8939.6 was excluded by GBA and only m/z 4289.5 was selected as the 4th feature, resulting in higher AUCs than *t* test alone.

Similar observations were also made for the features ranked by *t* test as 6th-9th and 13th-16th. For example, m/z 3890.6 and 7936.4 were mainly ranked at the 7th and 12th-14th positions respectively, but these features were found to be highly associated with m/z 7766.9, which is the top ranked feature (Table 5.2). m/z 7936, which was mainly ranked at the 13th or 15th positions, was also significantly correlated with m/z 3164.1, which was ranked 6th.

Table 5.2: The features selected by *t* test alone from the peaks identified by preprocessing during the 10-fold cross-validation. The first column represents the mass to charge ratios (m/z) of the features, and the second column contains their mean *t* values calculated from the 10-fold cross-validation. The ranks and numbers of occurrences of the features are shown next. For example, m/z 7766.9 in the first row has a mean *t* value of -0.00313 and was ranked #1 by *t* test alone six times (*i.e.*, 1(6)) and #2 four times (*i.e.*, 2(4)) during the 10-fold cross-validation. The features are sorted by their absolute mean *t* values.

m/z (Da)	Mean *t* value*	Rank (Frequency)				
7766.9	-0.00313	1(6)	2(4)			
4443.8	0.00265	1(2)	2(6)	3(2)		
2026.0	0.00201	1(2)	3(4)	4(2)	5(1)	
4289.5	0.00145	3(4)	4(5)	5(1)		
8939.6	0.00116	4(3)	5(7)			
3164.1	0.00084	5(1)	6(6)	7(1)	8(2)	
3890.6	-0.00077	6(1)	7(5)	8(2)	9(2)	
2232.6	0.00075	6(3)	7(3)	8(2)	9(1)	10(1)
4146.3	0.00067	7(1)	8(3)	9(5)	10(1)	
5911.2	0.00057	8(1)	9(2)	10(5)	11(1)	13(1)
3435.3	0.00051	10(3)	11(7)			
6986.8	0.00042	11(1)	12(6)	14(3)		
7936.4	-0.00038	12(3)	13(2)	14(4)	16(1)	
2745.0	0.00032	12(1)	13(4)	15(1)	16(2)	
2088.4	0.00030	13(2)	14(1)	15(4)	16(1)	
2795.2	0.00026	14(1)	15(2)	16(1)	19(1)	20(2)
7846.5	-0.00026	15(1)	16(2)	17(3)	18(1)	19(1)
6856.8	0.00025	15(1)	16(3)	17(1)	18(1)	20(2)
2304.3	-0.00024	15(1)	17(2)	18(3)	19(1)	
4747.8	0.00023	17(1)	18(1)	19(2)	20(3)	
8607.1	-0.00021	17(2)	18(2)	19(2)	20(1)	

* The mean t value of a feature was the average of its *t* values from the 10-fold cross-validation.

Table 5.3: The clusters of the features ranked as top 20 by *t* test during the 10-fold cross-validation. The clusters are sorted by their representative features' absolute mean *t* values.

Cluster	Feature (m/z)			
1	3890.6	7766.9*	7846.5	7936.4
2	4443.8*	4747.8	6986.8	8939.6
3	2026.0*			
4	4289.5*	6856.8		
5	3164.1*	3435.3	4146.3	
6	2088.4	2232.6*		
7	5911.2*			
8	2745.0*			
9	2304.3*			
10	2795.2*			
11	8607.1*			

* The feature with the highest discrimination ability in the cluster.

Another interesting finding about GBA is that it can work as a deconvolution algorithm for multiply charged proteins. Although, in general, molecules are singly charged by the MALDI (SELDI) ionizing process, sometimes molecules are multiply charged. If a molecule is multiply charged, its corresponding peak in a mass spectrum occurs approximately at the ratio of its molecular weight and charge amount. In general, it is impossible to know whether two peaks with multiple m/z values originate from the same molecule or not without prior information on the analyte. However, they are highly likely to represent the same protein if two peaks with multiply related m/z values have a strong correlation, which can be revealed by GBA. In Cluster 1 of Table 5.3, m/z 7766.9 is approximately twice m/z 3890.6 and these two features are highly correlated (correlation of 0.97). Similarly, m/z 8939.6 and m/z 4443.8 are also strongly

correlated (correlation of 0.85). Based on these observations, it is highly probable that m/z 7766.9 and m/z 3890.6 actually represent one protein, and likewise m/z 8939.6 and 4443.8 represent a single protein.

In summary, there are strong correlations between the features selected solely by *t* test, and these correlations are the major reason for performance degradation when features are selected. In contrast, when *t* test is used with GBA, the classification performance does not decrease as more features are selected; it remains steady since GBA avoids selecting features correlated strongly with other discriminant features. This observation is validated by the fact that the average distance of the features selected by *t* test with GBA was kept as high as 0.77 (correlation of 0.23) while some of the features selected by *t* test alone have high correlations with each other. The largest AUC of GBA was obtained when 16 features were used (≈ 0.80). However, I favor the simplest model for a given level of performance. This is often referred to as Occam's razor in machine learning. Thus, the top four features can be taken as potential biomarkers. In obtaining these four features as biomarker candidates, GBA eliminated a redundant feature from the list. Moreover, GBA can also provide a clue for identifying the features from multiply charged molecules.

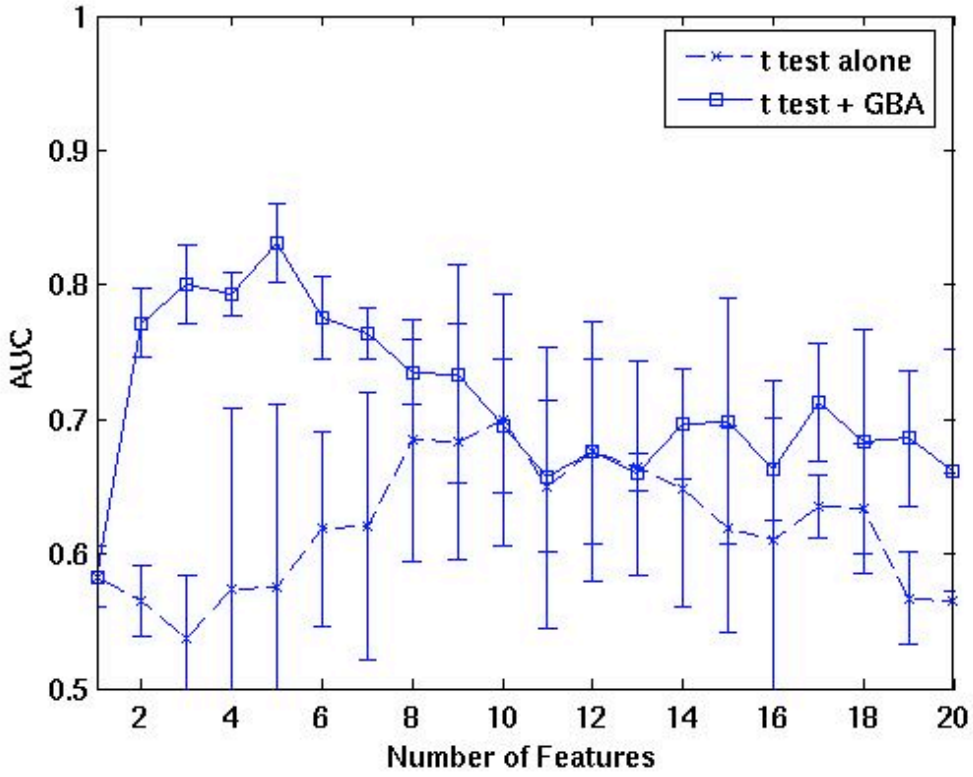

Figure 5.5: The AUCs of the logistic classifiers trained on the features selected by two sample *t* test alone and those by *t* test with GBA with the normalized mass spectra. The error bars represent the standard deviations obtained from three different 10-fold cross-validation runs. The comparison of *t* test alone and *t* test with GBA was repeated with the number of selected features varying from 1 to 20 (horizontal axis). *t* test with GBA is better than or comparable to *t* test alone. In particular, it is interesting that GBA can be used as an alternative to peak detection when applied to the normalized data since it successfully detected groups of features that represent specific proteins.

5.3.2 Two sample *t* test vs. *t* test with GBA on normalized mass spectra

The benefits of GBA were clearly apparent when it was applied to the normalized mass spectra without peak detection. As in the analysis of the peaks identified by preprocessing, I also performed three different 10-fold cross-validations to obtain the standard deviations of the AUCs. As can be seen in Figure 5.5, the AUCs were typically higher for *t* test with GBA than for *t* test alone except for when 10-13 features were used. The average AUC of GBA reached 0.84 when five features were used while the AUC of *t* test alone remained at 0.57. After the AUC of *t* test with GBA reached the maximum value, it continually decreased. The AUC of *t* test alone arrived at its maximum (≈ 0.70) until 10 features were selected and then gradually decreased below the AUC of *t* test with GBA. This observation means that the first 5-6 features selected by *t* test alone were highly correlated, but the features selected next are less correlated with the previous ones, resulting in the increasing AUC of *t* test alone until 10 features were included.

Table 5.4 shows the ranks of the top 20 features by *t* test alone. As can be seen in this table, it was observed that the features can be grouped by their m/z values. For example, the first eight features have m/z values in the range of from 2022.7 to 2028.4. Because the normalized mass spectra were directly used without peak detection, a protein is represented by several features with similar m/z values due to the finite resolution of the mass spectrometer and the isotopes of the protein. In general, peak detection algorithms are used in order to extract representative features that summarize the information of the proteins in the sample based on certain criteria such as signal to noise ratio. Moreover, since features with similar m/z values tend to be strongly correlated, peak detection algorithms actually select the representative features from the groups of

features with high correlations and similar m/z values. This implies that GBA can be used as an alternative to peak detection.

Table 5.5 shows the feature clusters containing the features that were ranked by *t* test alone as 1st-20th. As can be seen in Table 5.5, these features were clustered into four groups and the features in each group have similar m/z values, which implies that only four features among those selected by *t* test alone are independent, and this redundancy of the features lowered the AUC. It was observed that the distances within the clusters are very low (0-0.09) while the distances between the clusters range from 0.41 to 0.90. This observation shows that GBA can correctly detect the groups of features representing the same proteins and select features with high discrimination abilities when it is applied to raw mass spectra. Interestingly, these feature clusters correspond to those obtained in the peaks identified by preprocessing. Cluster 1 includes the features from m/z 7765.3 to m/z 7778.3, which correspond to m/z 7766.9 of Cluster 1 in Table 5.3. Similarly, Clusters 2, 3, and 4 are directly associated with the representative features of Clusters 2, 3, and 4 (m/z 4443.8, 2026.0, and 4289.5, respectively) in Table 5.3. This result also supports that GBA does peak detection implicitly and assists in identifying discriminant ones from the detected features.

The use of GBA for normalized mass spectra made it possible to select features that are independent as well as discriminant. The effect of GBA was clearer for the case in which raw mass spectra were directly used than for the case in which peaks were identified by the preprocessing because GBA successfully grouped the groups of features that may represent the same proteins based on only correlations among the features, and selected the most discriminant representative features from the groups. This experimental result also indicates that GBA performs peak detection without heuristic parameter determination as required by other peak detection algorithms.

Table 5.4: The features selected by two sample *t* test alone from the normalized mass spectra during the 10-fold cross-validation. The first column represents the mass to charge ratios (m/z) of the features, and the second column contains their mean *t* values calculated from the 10-fold cross-validation. The ranks and numbers of occurrences of the features are shown next. The features are also sorted by their absolute mean *t* values.

m/z (Da)	Mean *t* value*	Rank (Frequency)					
7771.8	-0.00302	1(6)	2(1)	4(1)	7(1)	9(1)	
7770.1	-0.00294	2(4)	3(2)	4(1)	7(1)	8(1)	13(1)
7773.4	-0.00294	2(2)	3(4)	6(2)	8(1)	9(1)	
7775.0	-0.00273	4(2)	5(3)	7(1)	9(1)	10(1)	11(1)
7768.5	-0.00271	4(3)	5(1)	7(1)	8(1)	9(1)	10(2)
4443.8	0.00264	1(2)	4(1)	5(1)	6(2)	7(1)	14(1)
4442.6	0.00262	2(1)	3(1)	5(1)	6(1)	7(1)	8(1)
4445.0	0.00249	3(1)	6(1)	8(2)	9(1)	10(1)	12(2)
7776.6	-0.00243	6(1)	8(1)	9(2)	10(2)	13(1)	14(1)
4441.3	0.00237	4(1)	9(2)	11(3)	15(1)	19(1)	20(1)
7766.9	-0.00236	6(1)	10(2)	11(2)	12(1)	14(1)	15(1)
4446.2	0.00220	5(1)	11(1)	12(2)	13(1)	16(1)	17(1)
7778.3	-0.00208	12(1)	13(2)	14(2)	18(1)	19(2)	
2025.1	0.00205	1(2)	5(1)	7(1)	8(1)	14(1)	
4440.1	0.00200	6(1)	13(2)	15(2)	18(1)		
2026.0	0.00198	3(1)	4(1)	7(1)	8(1)	9(1)	16(1)
7765.3	-0.00194	15(1)	16(1)	18(1)			
2024.3	0.00194	2(2)	7(1)	11(1)	14(1)	20(1)	
2026.8	0.00187	5(1)	8(1)	10(1)	13(2)	19(1)	
2023.5	0.00185	5(1)	12(1)	17(1)	18(1)		
4447.5	0.00183	7(1)	15(1)	16(2)			
7779.9	-0.00173	17(3)	18(1)				
2027.6	0.00166	12(3)	16(1)	19(1)			
2022.7	0.00161	6(1)	11(1)	17(1)			
4438.9	0.00159	11(1)	17(1)	18(1)			
4448.7	0.00147	12(1)	19(1)				
2028.4	0.00145	18(1)	19(1)	20(1)			
4285.9	0.00144	16(1)	18(1)	19(1)	20(1)		
4287.1	0.00143	15(1)	19(1)	20(1)			

* The mean t value of a feature was the average of its *t* values from the 10-fold cross-validation.

Table 5.5: The clusters of the features ranked as top 20 by *t* test from the normalized mass spectra during the 10-fold cross-validation. The clustering results were very consistent throughout the cross-validation. All of the features ranked within 20 were grouped into 4 clusters. The clusters are also sorted by their representative features' absolute mean *t* values.

Cluster	m/z (Da)								
1	7765.3	7766.9	7768.5	7770.1	7771.8*	7773.4	7775	7776.6	7778.3
2	4438.9	4440.1	4441.3	4442.6	4443.8*	4445.0	4446.2	4447.5	4448.7
3	4285.9*	4287.1							
4	2022.7	2023.5	2024.3	2025.1*	2026.0	2026.8	2027.6	2028.4	

* The feature with the highest discrimination ability in the cluster.

Table 5.6: The features frequently ranked from 1 to 5 by t test with GBA from the peaks identified by preprocessing and the normalized spectra are compared with those reported in Pusztai *et al.*'s study. It is observed that the features selected from the peaks identified by preprocessing and those from the normalized spectra have very similar mass to charge ratios. This implies that GBA can be used as an alternative to peak detection. Two features were commonly identified by *t* test with GBA and Pusztai *et al.*'s method (m/z 4444 and 3165). According to Table 5.3, the m/z 4444 is highly correlated with m/z 8940 (≈ 0.85) and the m/z is almost half of 8940. It is highly probable that m/z 4444 and m/z 8940 represent multiply charged states of the biologically same protein.

t test with GBA		Pusztai *et al*. Cancer 2004 (Da)
Peaks identified by preprocessing (Da)	Normalized spectra (Da)	
2026.0	2025.1	
3164.1	*3162.0*	*3165*
		3440
		4115
4289.5	4285.9	
4443.8	*4443.8*	*4444*
7766.9	7771.8	
		8940

5.3.3 Comparison with Pusztai *et al.*'s finding

The features selected by *t* test with GBA from the peaks identified by preprocessing and those from the normalized mass spectra were compared with the features reported by Pusztai *et al* [111]. The features in the first two columns of Table 5.6 are the top five features selected by *t* test with GBA from the peaks identified by preprocessing and the normalized mass spectra respectively, and the features in the last column are the five features identified by Pusztai *et al*. As can be seen in Table 5.6, the m/z values of the features selected from the peaks identified by preprocessing agreed with those selected from the normalized mass spectra. This indicates that GBA is able to extract non-redundant features that represent the proteins in the analyte, so it can be used as an alternative to peak detection.

Two of the features selected by *t* test with GBA (m/z 4444 and m/z 3165) are in the feature list reported by Pusztai *et al*. They had also used the same *t* test; however, Pusztai *et al*. normalized the mass spectra with respect to the total ion current after the 1000^{th} time tick [111]; thus, their list of features is not exactly the same as ours. Moreover, my GBA analysis implies that the feature m/z 4444 corresponds to doubly charged molecules of m/z 8940 because they are highly correlated (correlation of 0.85).

5.4 DISCUSSIONS AND CONCLUSION

To achieve accurate classification performance, features must be both discriminant and independent. In general, feature selection methods, especially filter methods, do not take into account the relationships between features, which can lead to the degradation of overall discrimination ability. On the other hand, wrapper-based feature selection methods search for a best set of features by optimizing the group discrimination ability. However, the involvement of a search algorithm in feature

selection can increase the computational load (*e.g.*, genetic algorithms) or result in bias during searching (*e.g.*, forward selection or backward elimination). Therefore, feature selection based on a wrapper method can take a long time and may converge to solution that is only locally optimal. In comparison with these conventional feature selection methods, GBA successfully eliminates the redundancy among features by clustering closely correlated features, which enables the selection of more discriminant and independent features in a moderate amount of time.

When there are strong correlations between discriminant features, using a filter method with GBA is more beneficial than using a filter only because GBA eliminates the redundancy between discriminant features, maximizing the discrimination ability of the selected feature. When there are not strong correlations between discriminant features or when there are strong correlations only between non-discriminant features, using a filter method with GBA is comparable to using a filter alone and two methods would end up with similar features.

In addition, GBA makes it possible to identify features that represent multiply charged molecules of the same protein by elucidating the relationships between features. The experimental result of direct GBA application to the normalized mass spectra without peak detection demonstrates that GBA can serve as an alternative to peak detection and peak alignment because multiple features that originate from the same protein molecules have strong correlations and similar m/z values. Most conventional peak detection and alignment methods reduce feature redundancy with only a local scope, and parameters relating to resolution, mass accuracy, and signal to noise ratio must be set heuristically by the user. GBA eliminates feature redundancy based on a global correlation measure and it does not require such parameter choices. However, it should be noted that the computational running time increases quadratically with the number of

features. Thus, the runtime of GBA can be slow when a large number of features are analyzed.

The next step to identifying biomarker candidates would be to prove whether the features identified are biologically meaningful. In other words, the proteins corresponding to the selected peaks should be identified. Pusztai *et al.* searched ExPASy [213] to identify proteins of approximately the m/z identified in their analysis [111]. I also attempted to identify proteins with the similar mass values of my findings. However, it is difficult to identify proteins using only their mass information because even the simplest search, TagIdent, requires the mass range and isoelectric point (PI) range as minimum parameters. Obviously, different combinations of parameter values can lead to different search results for a given mass value. In my search, I arbitrarily set the relative mass range to 1% and the PI range from 4.5 to 8.5 [214], which is believed to cover most of the PI ranges of human proteins. For m/z 7766.9, a TagIdent search returned eight possible matches (Table 5.7). Some of the matched proteins are related with diseases. For example, Small inducible cytokine A3-like 1 and Small inducible cytokine A3 are known to work as an inhibitor of HIV-I virus [215]. P8 MTCP-1 is also related with a disease; it was reported to be overexpressed in T-cell leukemia. None of these proteins also do have any known direct association with breast cancer, but m/z 7766.9 seems to reflect the immune responses of the body with respect to breast cancer. I performed similar searches for the other proteins presented in Table 5.6, but I could not identify proteins that are known to be directly related with breast cancer. For more accurate results, protein/peptide identification through immunoassay or tandem mass spectrometry would be necessary.

Table 5.7: The eight possible proteins that match to m/z 7766.9 based on the TagIdent search results. The relative mass error was set to 1%, which is believed as the typical mass error rate of SELDI TOF, and the PI was also arbitrarily set to 4.5-8.5.

m/z (Da)	Possible Proteins
7766.9	Small inducible cytokine A14
	Small inducible cytokine A3-like 1
	Small inducible cytokine A3
	Small inducible cytokine A4
	High affinity immunoglobulin epsilon receptor gamma-subunit
	Platelet factor 4 variant precursor
	P8 MTCP-1
	Protein UNQ655/PRO1286

In conclusion, this chapter presents a feature selection algorithm that searches for discriminant and independent features through feature clustering. In my study, GBA almost always yielded better results than *t* test alone when GBA was tested on the breast cancer SELDI TOF data set. Even though GBA was designed for proteomic profiling, GBA can be extended to other types of data that require extensive feature selection. Ordinal features can also be handled by GBA if the rank correlation is used to estimate the correlations between features. GBA also has the potential to reduce the biochemical/ computational burdens by means of suggesting non-redundant candidates of potential disease biomarkers from a complex protein samples like SELDI TOF mass spectra. In addition, GBA showed potential as a feature reduction technique without peak detection by removing peaks from related protein species.

Chapter 6: DISCUSSIONS AND CONCLUSION

6.1 SUMMARY OF WORK

Proteomic profiling by MALDI TOF MS is a very powerful method for identifying potential biomarkers for diagnosis and prognosis of disease. However, since MALDI TOF suffers from several types of noise, the subtle but pathologically useful information in the mass spectra can be difficult to discern. On the other hand, due to the notoriously complex nature of protein profiles by MALDI TOF MS, methods are needed for selecting the most informative peaks from the spectra.

In an effort to reducing noise in MALDI TOF MS, an extensive noise model was proposed and each noise component was separated and analyzed according to the model. In Chapter 3, I hypothesized that noise in MALDI TOF MS is composed of three major components of noise: noise from instrumentation, noise from random ion motion, and noise from chemical impurities in the sample (chemical noise). Noise from instrumentation was analyzed using parametric power spectral density estimation. In this dissertation, I employed the Burg algorithm for multiple segments developed by de Waele and Broersen [107] in order to increase the accuracy of the estimation. The experimental results revealed that noise from instrumentation is a mixture of several types of physical and electrical noise such as thermal noise, $1/f$ noise, and periodic interferences from the internal/external circuits of the mass spectrometer. My simulation study also showed that noise from instrumentation might not make a strong impact on the quality of mass spectra. However, unlike noise from instrumentation, chemical noise seriously affects mass spectra by adding a monotonically decreasing baseline and high frequency noise, particularly to the low mass region of the spectra. Therefore, in Chapter 4, I characterized chemical noise and developed a wavelet-based denoising algorithm.

Chemical noise in MALDI is mainly caused by matrix material in the sample, which is small organic substance such as sinapinic acid. The matrix material helps proteins/peptides to be ionized without thermal damage by the laser. However, since the matrix molecules are also ionized together with analytes by the laser, they also produce undesired interferences with the signal of proteins/peptides in the sample. Chemical noise is non-stationary and non-white, characterized by its monotonically decreasing baseline and variance over time. I developed a denoising method for estimating adaptive thresholds using multiple realizations of chemical noise and a baseline correction algorithm for eliminating the baseline in the wavelet domain. The comparison of my algorithms with other conventional denoising and baseline correction algorithms showed that my algorithms are superior to traditional methods.

In machine learning, ideal features are discriminant and independent. My novel feature selection algorithm, guilt-by-association (GBA) feature selection, can make it possible to do this through a correlation-based similarity measure and clustering algorithm. GBA groups highly similar features together and selects the representative features in conjunction with a regular filter method such as the two sample t test. The efficacy of GBA was investigated using a simulation and a real breast cancer SELDI TOF data set. The experimental results showed that GBA in combination with the t test improved classification performance. Moreover, GBA demonstrated its potential as an alternative to conventional peak detection and alignment algorithms by successfully identifying local peak groups. It also deconvolved multiply charged states of the same proteins.

In this dissertation, novel methods for denoising and feature selection were presented. The methods have been developed to handle complex proteomic profiles by MALDI TOF MS for potential biomarker identification, and have shown their

tremendous potential for removing artifacts and extracting the key information for characterizing the pathological state of a disease from raw data. In the following section, several suggestions for future studies will be made from an engineering perspective.

6.2 SUGGESTIONS FOR FUTURE STUDIES

As described in Chapter 2, biomarker identification utilizing mass spectrometry generally consists of five steps: preprocessing, feature extraction, feature selection, classification, and evaluation. When this research area began to attract scientific interests, many computer scientists or computational biologists put more weight on classification. They attempted to find subtle molecular "patterns" from mass spectra by almost blindly applying sophisticated machine learning algorithms and reported that their experimental results looked promising. However, subsequent studies questioned about the efficacy of these initial experiments, pointing out that the "patterns" identified by the machine learning algorithms could be simply due to chance differences, between the diseased and healthy groups. Since then, preprocessing, feature extraction, and feature selection have attracted more attention because these methods can improve the reliability of classification. No matter how effective a machine learning algorithm is used, it cannot be expected to have statistically reliable results from poorly preprocessed data or coarsely extracted and selected features. However, despite of the importance of these research areas, the studies regarding preprocessing, feature extraction, and feature selection for mass spectrometry are still in their infancy.

In particular, noise analysis still remains underexplored. In previous studies, *ad hoc* methods such as use of a moving average window have been employed to reduce noise without any efforts at first statistically characterizing noise. Therefore, it is necessary for researchers with electrical engineering backgrounds, particularly stochastic

signal processing, to join this research area for rigorous analysis and reduction of noise in mass spectrometry. The use of wavelet analysis or wavelet packet analysis for denoising mass spectra is particularly promising (Chapter 4). Research on adjusting and applying these powerful signal processing techniques for noise analysis mass spectrometry is in the early stages.

Since biomarker identification using mass spectrometry generally includes comparison of multiple mass spectra, normalization must be applied to make the mass spectra comparable to each other. The importance of normalization in microarray data analysis has been demonstrated by several studies [216, 217]. However, up to the best of my knowledge, there have no studies that extensively compare a variety of normalization methods in mass spectrometry data analysis. Studies regarding normalization will also make significant impacts on the biomarker identification research area.

Unlike typical machine learning tasks, only a moderately small number of samples are available while a huge number of features are produced by mass spectrometry (Chapter 2). Thus, the learning algorithm suffers from the *curse of dimensionality* (Chapter 2 and 5). It would be beneficial to develop algorithms that can efficiently reduce the dimensionality while keeping statistically important features in the reduced feature set.

Finally, prompt biological evaluation of findings obtained from feature selection or classification are needed. For example, when potential biomarkers are proposed, these candidates should be promptly validated using protein/peptide identification through immunoassay or tandem mass spectrometry. In Chapter 5, I performed a very simple peptide/protein search using TagIdent provided by ExPASy, but it did not help to clearly identify the candidates. Thus, more sophisticated biochemical validation must follow the

types of analysis presented in this dissertation. Increased collaboration between engineers and biochemists will be needed to achieve this goal.

Bibliography

[1] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, C. Smigal, and M. J. Thun, "Cancer Statistics, 2006," *CA: A Cancer Journal for Clinicians*, vol. 56, pp. 106-130, 2006.

[2] "Cancer Facts and Figures 2006," American Cancer Society 2006.

[3] R. Etzioni, N. Urban, S. Ramsey, M. McIntosh, S. Schwartz, B. Reid, J. Radich, G. Anderson, and L. Hartwell, "The case for early detection," *Nature Reviews. Cancer*, vol. 3, pp. 243-52, 2003.

[4] M. T. Fahey, L. Irwig, and P. Macaskill, "Meta-analysis of Pap test accuracy.[see comment]," *American Journal of Epidemiology*, vol. 141, pp. 680-9, 1995.

[5] C. H. Lee, "Screening mammography: proven benefit, continued controversy," *Radiologic clinics of North America*, vol. 40, pp. 395-407, 2002.

[6] B. B. Green and S. H. Taplin, "Breast Cancer Screening Controversies," *J Am Board Fam Pract*, vol. 16, pp. 233-241, 2003.

[7] D. B. Kopans, "The positive predictive value of mammography," *AJR. American Journal of Roentgenology*, vol. 158, pp. 521-6, 1992.

[8] A. M. Knutzen and J. J. Gisvold, "Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions," *Mayo Clinic Proceedings.*, vol. 68, pp. 454-60, 1993.

[9] L. L. Humphrey, M. Helfand, B. K. S. Chan, and S. H. Woolf, "Breast Cancer Screening: A Summary of the Evidence for the U.S. Preventive Services Task Force," *Ann Intern Med*, vol. 137, pp. 347-360, 2002.

[10] J. M. E. Walsh and J. P. Terdiman, "Colorectal Cancer Screening: Clinical Applications," *JAMA*, vol. 289, pp. 1297-1302, 2003.

[11] J. M. E. Walsh and J. P. Terdiman, "Colorectal Cancer Screening: Scientific Review," *JAMA*, vol. 289, pp. 1288-1296, 2003.

[12] G. Rennert, H. S. Rennert, E. Miron, and Y. Peterburg, "Population Colorectal Cancer Screening with Fecal Occult Blood Test," *Cancer Epidemiol Biomarkers Prev*, vol. 10, pp. 1165-1168, 2001.

[13] M. Pignone, M. Rich, S. M. Teutsch, A. O. Berg, and K. N. Lohr, "Screening for Colorectal Cancer in Adults at Average Risk: A Summary of the Evidence for the U.S. Preventive Services Task Force," *Ann Intern Med*, vol. 137, pp. 132-141, 2002.

[14] S. W. Vernon, "Participation in colorectal cancer screening: a review," *Journal of the National Cancer Institute*, vol. 89, pp. 1406-22, 1997.

[15] M. E. Peek and J. H. Han, "Disparities in screening mammography. Current status, interventions and implications," *Journal of General Internal Medicine*, vol. 19, pp. 184-94, 2004.

[16] M. K. Brawer, "Prostate-specific antigen: current status," *CA Cancer J Clin*, vol. 49, pp. 264-281, 1999.

[17] P. R. Srinivas, S. Srivastava, S. Hanash, and G. L. Wright, Jr, "Proteomics in Early Detection of Cancer," *Clin Chem*, vol. 47, pp. 1901-1911, 2001.

[18] L. A. Liotta, M. Ferrari, and E. Petricoin, "Written in Blood," *Nature*, vol. 425, pp. 905, 2003.

[19] J. D. Wulfkuhle, L. A. Liotta, and E. F. Petricoin, "Proteomic applications for the early detection of cancer," *Nature Reviews. Cancer*, vol. 3, pp. 267-75, 2003.

[20] W. Pusch, M. T. Flocco, S. M. Leung, H. Thiele, and M. Kostrzewa, "Mass spectrometry-based clinical proteomics," *Pharmacogenomics*, vol. 4, pp. 463-76, 2003.

[21] R. Woolas, F. Xu, I. Jacobs, Y. Yu, L. Daly, A. Berchuck, J. Soper, D. Clarke- Pearson, D. Oram, and R. Bast, Jr, "Elevation of multiple serum markers in patients with stage I ovarian cancer," *J Natl Cancer Inst*, vol. 85, pp. 1748-1751, 1993.

[22] M. Verma, G. L. Wright, Jr., S. M. Hanash, R. Gopal-Srivastava, and S. Srivastava, "Proteomic approaches within the NCI early detection research network for the discovery and identification of cancer biomarkers," *Annals of the New York Academy of Sciences*, vol. 945, pp. 103-15, 2001.

[23] A. Madi, T. Pusztahelyi, M. Punyiczki, and L. Fesus, "The biology of the post-genomic era: the proteomics," *Acta Biologica Hungarica*, vol. 54, pp. 1-14, 2003.

[24] A. Abbott, "A post-genomic challenge: learning to read patterns of protein synthesis," *Nature*, vol. 402, pp. 715-20, 1999.

[25] A. J. Rai and D. W. CHan, "Cancer Proteomics: Serum Diagnostics for Tumor Marker Discovery," *Ann NY Acad Sci*, vol. 1022, pp. 286-294, 2004.

[26] K. D. Rodland, "Proteomics and cancer diagnosis: the potential of mass spectrometry," *Clinical Biochemistry*, vol. 37, pp. 579-583, 2004.

[27] E. E. Petricoin, C. P. Paweletz, and L. A. Liotta, "Clinical applications of proteomics: proteomic pattern diagnostics.," *Journal of Mammary Gland Biology & Neoplasia*, vol. 7, pp. 433-440, 2002.

[28] E. F. Petricoin, D. A. Fishman, T. P. Conrads, T. D. Veenstra, and L. A. Liotta, "Lessons from Kitty Hawk: From feasibility to routine clinical use for the field of proteomic pattern diagnostics," *Proteomics*, vol. 4, pp. 2357-2360, 2004.

[29] K. P. Rosenblatt, P. Bryant-Greenwood, J. K. Killian, A. Mehta, D. Geho, V. Espina, E. F. Petricoin, and L. A. Liotta, "Serum Proteomics in Cancer Diagnosis and Management," *Annual Review of Medicine*, vol. 55, pp. 97-112, 2004.

[30] T. P. Conrads, M. Zhou, E. F. Petricoin, 3rd, L. Liotta, and T. D. Veenstra, "Cancer diagnosis using proteomic patterns," *Expert Review of Molecular Diagnostics*, vol. 3, pp. 411-20, 2003.

[31] R. C. Krieg, C. P. Paweletz, L. A. Liotta, and E. F. Petricoin, 3rd, "Clinical proteomics for cancer biomarker discovery and therapeutic targeting," *Technology in Cancer Research & Treatment*, vol. 1, pp. 263-72, 2002.

[32] J. Li, H. Liu, S. Ng, and L. Wong "Discovery of significant rules for classifying cancer diagnosis data," *Bioinformatics*, vol. 19, pp. 93-102, 2003.

[33] G. Alexe, S. Alexe, L. A. Liotta, E. F. Petricoin, M. Reiss, and P. L. Hammer, "Ovarian cancer detection by logical analysis of proteomic data," *Proteomics*, vol. 4, pp. 766-783, 2004.

[34] D. J. Johann, Jr., M. D. McGuigan, S. Tomov, V. A. Fusaro, S. Ross, T. P. Conrads, T. D. Veenstra, D. A. Fishman, G. R. Whiteley, E. F. Petricoin, and L. A. Liotta, "Novel approaches to visualization and data mining reveals diagnostic information in the low amplitude region of serum mass spectra from ovarian cancer patients," *Disease Markers*, vol. 19, pp. 197-207, 2003-2004.

[35] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm, and J. S. Kovach, "Detection of cancer-specific markers amid massive mass spectral data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 14666-71, 2003.

[36] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet.*, vol. 359, pp. 572-7, 2002.

[37] K. R. Kozak, M. W. Amneus, S. M. Pusey, F. Su, M. N. Luong, S. A. Luong, S. T. Reddy, and R. Farias-Eisner, "Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: Potential use in diagnosis and prognosis," *PNAS*, vol. 100, pp. 12343-12348, 2003.

[38] J. M. Sorace and M. Zhan, "A data review and re-assessment of ovarian cancer serum proteomic profiling," *BMC bioinformatics*, vol. 4, 2003.

[39] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, pp. 1636-1643, 2003.

[40] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, and R. A. Clark, "Data mining techniques for cancer detection using serum proteomic profiling," *Artificial Intelligence in Medicine*, vol. 32, pp. 71-83, 2004.

[41] R. R. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics*, vol. 19, pp. 1484-1491, 2003.

[42] N. Jeffries, "Performance of a genetic algorithm for mass spectrometry proteomics," *BMC Bioinformatics*, vol. 5, pp. 180, 2004.

[43] T. P. Conrads, V. A. Fusaro, S. Ross, D. Johann, V. Rajapakse, B. A. Hitt, S. M. Steinberg, E. C. Kohn, D. A. Fishman, G. Whitely, J. C. Barrett, L. A. Liotta, Petricoin Ef, 3rd, and T. D. Veenstra, "High-resolution serum proteomic features for ovarian cancer detection," *Endocr Relat Cancer*, vol. 11, pp. 163-178, 2004.

[44] I. Levner, "Feature selection and nearest centroid classification for protein mass spectrometry," *BMC Bioinformatics*, vol. 6, pp. 68, 2005.

[45] J. Yu and X. W. Chen, "Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data," *Bioinformatics*, vol. 21 Suppl 1, pp. i487-i494, 2005.

[46] J. S. Yu, S. Ongarello, R. Fiedler, X. W. Chen, G. Toffolo, C. Cobelli, and Z. Trajanoski, "Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data," *Bioinformatics*, vol. 21, pp. 2200-2209, 2005.

[47] L. H. Cazares, B. L. Adam, M. D. Ward, S. Nasim, P. F. Schellhammer, O. J. Semmes, and G. L. Wright, Jr., "Normal, benign, preneoplastic, and malignant prostate cells have distinct protein expression profiles resolved by surface enhanced laser desorption/ionization mass spectrometry," *Clinical Cancer Research*, vol. 8, pp. 2541-52, 2002.

[48] R. H. Lilien, H. Farid, and B. R. Donald, "Probabilistic Disease Classification of Expression-Dependent Proteomic Data from Mass Spectrometry of Human Serum," *Journal of Computational Biology*, vol. 10, pp. 925-946, 2003.

[49] M. Wagner, D. Naik, A. Pothen, S. Kasukurti, R. Devineni, B.-L. Adam, O. J. Semmes, and G. Wright, "Computational protein biomarker prediction: a case study for prostate cancer," *BMC Bioinformatics*, vol. 5, pp. 26, 2004.

[50] J. Li, N. White, Z. Zhang, J. Rosenzweig, L. A. Mangold, A. W. Partin, and D. W. Chan, "Detection of Prostate Cancer Using Serum Proteomics Pattern in a Histologically Confirmed Population," *Journal of Urology*, vol. 171, pp. 1782-1787, 2004.

[51] B. L. Adam, Y. Qu, J. W. Davis, M. D. Ward, M. A. Clements, L. H. Cazares, O. J. Semmes, P. F. Schellhammer, Y. Yasui, Z. Feng, and G. L. Wright, Jr., "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men," *Cancer Research*, vol. 62, pp. 3609-14, 2002.

[52] Y. Yasui, M. Pepe, M. L. Thompson, B. L. Adam, G. L. Wright, Jr., Y. Qu, J. D. Potter, M. Winget, M. Thornquist, and Z. Feng, "A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection," *Biostatistics*, vol. 4, pp. 449-463, 2003.

[53] Y. Qu, B.-L. Adam, Y. Yasui, M. D. Ward, L. H. Cazares, P. F. Schellhammer, Z. Feng, O. J. Semmes, and G. L. Wright, Jr., "Boosted Decision Tree Analysis of Surface-enhanced Laser Desorption/Ionization Mass Spectral Serum Profiles Discriminates Prostate Cancer from Noncancer Patients," *Clin Chem*, vol. 48, pp. 1835-1843, 2002.

[54] Y. Qu, B. L. Adam, M. Thornquist, J. D. Potter, M. L. Thompson, Y. Yasui, J. Davis, P. F. Schellhammer, L. Cazares, M. Clements, G. L. Wright, Jr., and Z. Feng, "Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data," *Biometrics*, vol. 59, pp. 143-51, 2003.

[55] J. H. Stone, V. N. Rajapakse, G. S. Hoffman, U. Specks, P. A. Merkel, R. F. Spiera, J. C. Davis, E. W. St.Clair, J. McCune, S. Ross, B. A. Hitt, T. D. Veenstra, T. P. Conrads, L. A. Liotta, and E. F. P. III, "A serum proteomic approach to gauging the state of remission in Wegener's granulomatosis," *Arthritis & Rheumatism*, vol. 52, pp. 902-10, 2005.

[56] D. K. Ornstein, W. Rayford, V. A. Fusaro, T. P. Conrads, S. J. Ross, B. A. Hitt, W. W. Wiggins, T. D. Veenstra, L. A. Liotta, and E. F. Petricoin, 3rd, "Serum proteomic profiling can discriminate prostate cancer from benign prostates in men with total prostate specific antigen levels between 2.5 and 15.0 ng/ml," *Journal of Urology*, vol. 172, pp. 1302-5, 2004.

[57] G. Bhanot, G. Alexe, B. Venkataraghavan, and A. J. Levine, "A robust meta-classification strategy for cancer detection from MS data," *Proteomics*, vol. 6, pp. 592-604, 2006.

[58] J. Li, Z. Zhang, J. Rosenzweig, Y. Y. Wang, and D. W. Chan, "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer," *Clinical Chemistry*, vol. 48, pp. 1296-304, 2002.

[59] S. Becker, L. Cazares, P. Watson, H. Lynch, O. Semmes, R. Drake, and C. Laronga, "Surfaced-Enhanced Laser Desorption/Ionization Time-of-Flight(SELDI-TOF) Differentiation of Serum Protein Profiles of BRCA-1 and Sporadic Breast Cancer," *Annals of Surgical Oncology*, vol. 11, pp. 907-914, 2004.

[60] H. M. Kuerer, K. R. Coombes, J. N. Chen, L. Xiao, C. Clarke, H. Fritsche, S. Krishnamurthy, S. Marcy, M. C. Hung, and K. K. Hunt, "Association between ductal fluid proteomic expression profiles and the presence of lymph node metastases in women with breast cancer," *Surgery*, vol. 136, pp. 1061-1069, 2004.

[61] A. Vlahou, P. F. Schellhammer, and G. L. Wright, Jr., "Application of a Novel Protein Chip Mass Spectrometry Technology for the Identification of Bladder Cancer-Associated Biomarkers," *Advances in Experimental Medicine and Biology*, vol. 539A, pp. 47-60, 2003.

[62] A. Vlahou, A. Giannopoulos, B. W. Gregory, T. Manousakas, F. I. Kondylis, L. L. Wilson, P. F. Schellhammer, G. L. Wright, Jr., and O. J. Semmes, "Protein profiling in urine for the diagnosis of bladder cancer," *Clinical Chemistry*, vol. 50, pp. 1438-41, 2004.

[63] Q. Liu, B. Krishnapuram, P. Pratapa, X. Liao, A. Hartemink, and L. Carin, "Identification of Differentially Expressed Proteins Using MALDI-TOF Mass Spectra," presented at ASILOMAR Conference: Biological Aspects of Signal Processing, 2003.

[64] M. J. Campa, M. Z. Wang, B. Howard, M. C. Fitzgerald, and E. F. Patz, Jr., "Protein expression profiling identifies macrophage migration inhibitory factor and cyclophilin a as potential molecular targets in non-small cell lung cancer," *Cancer Res*, vol. 63, pp. 1652-6, 2003.

[65] T. A. Zhukov, R. A. Johnson, A. B. Cantor, R. A. Clark, and M. S. Tockman, "Discovery of distict protein profiles specific for lung tumors and pre-malignat lung lesions by SELDI mass spectrometry," *Lung Cancer*, vol. 40, pp. 267-79, 2003.

[66] M. Hilario, A. Kalousis, M. Muller, and C. Pellegrini, "Machine learning approaches to lung cancer prediction from mass spectra," *Proteomics*, vol. 3, pp. 1716-1719, 2003.

[67] P. Neville, P. Tan, G. Mann, and R. Wolfinger, "Generalizable mass spectrometry mining used to identify disease state biomarkers from blood serum," *Proteomics*, vol. 3, pp. 1710-1715, 2003.

[68] M. Wagner, D. Naik, and A. Pothen, "Protocols for disease classification from mass spectrometry data," *Proteomics*, vol. 3, pp. 1692-1698, 2003.

[69] D. Sidransky, R. Irizarry, J. A. Califano, X. Li, H. Ren, N. Benoit, and L. Mao, "Serum Protein MALDI Profiling to Distinguish Upper Aerodigestive Tract Cancer Patients From Control Subjects," *J Natl Cancer Inst*, vol. 95, pp. 1711-1717, 2003.

[70] K. Yanagisawa, B. J. Xu, P. P. Massion, P. H. Larsen, B. C. White, J. R. Roberts, M. Edgerton, A. Gonzalez, S. Nadaf, J. H. Moore, R. M. Caprioli, and D. P. Carbone, "Proteomic patterns of tumour subsets in non-small-cell lung cancer," *Lancet*, vol. 362, pp. 433-439, 2003.

[71] K. A. Baggerly, J. S. Morris, J. Wang, D. Gold, L. Xiao, and K. R. Coombes, "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples," *Proteomics*, vol. 3, pp. 1667-1672, 2003.

[72] M. K. Markey, G. D. Tourassi, and C. E. J. Floyd, "Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer," *Proteomics*, vol. 3, pp. 1678-1679, 2003.

[73] K. R. Lee, X. Lin, D. C. Park, and S. Eslava, "Megavariate data analysis of mass spectormetric proteomics data using latent varaible projection method.," *Proteomics*, vol. 3, pp. 1680-1686, 2003.

[74] D. J. Slotta, L. S. Heath, N. Ramakrishnan, R. Helm, and M. Potts, "Clustering mass spectrometry data using order statistics," *Proteomics*, vol. 3, pp. 1687-1691, 2003.

[75] P. V. Purohit and D. M. Rocke, "Discriminant models for high-throughput proteomics mass spectrometer data," *Proteomics*, vol. 3, pp. 1699-1703, 2003.

[76] J. W. Tatay, X. Feng, N. Sobczak, H. Jiang, C. Chen, R. Kirova, C. Struble, N. J. Wang, and P. J. Tonellato, "Multiple approaches to data-mining of proteomics data based on statistical and pattern classification methods," *Proteomics*, vol. 3, pp. 1704-1709, 2003.

[77] H. Zhu, C. Y. Yu, and H. Zhang, "Tree-based disease classification using protein data," *Proteomics.*, vol. 3, pp. 1673-7, 2003.

[78] M. Seike, T. Kondo, K. Fujii, T. Yamada, A. Gemma, S. Kudoh, and S. Hirohashi, "Proteomic signature of human cancer cells," *Proteomics*, vol. 4, pp. 2776-2778, 2004.

[79] T. C. W. Poon, T.-T. Yip, A. T. C. Chan, C. Yip, V. Yip, T. S. K. Mok, C. C. Y. Lee, T. W. T. Leung, S. K. W. Ho, and P. J. Johnson, "Comprehensive Proteomic Profiling Identifies Serum Proteomic Signatures for Detection of Hepatocellular Carcinoma and Its Subtypes," *Clin Chem*, vol. 49, pp. 752-760, 2003.

[80] S. Bhattacharyya, E. R. Siegel, G. M. Petersen, S. T. Chari, L. J. Suva, and R. S. Haun, "Diagnosis of Pancreatic Cancer Using Serum Proteomic Profiling," *Neoplasia*, vol. 6, pp. 674-686, 2004.

[81] A. Valerio, D. Basso, P. Fogar, M. Falconi, E. Greco, C. Bassi, R. Seraglia, M. Abu-Hilal, F. Navaglia, and C.-F. Zambon, "Maldi-TOF analysis of portal sera of pancreatic cancer patients: identification of diabetogenic and antidiabetogenic peptides," *Clinica Chimica Acta*, vol. 343, pp. 119-127, 2004.

[82] A. Valerio, D. Basso, S. Mazza, G. Baldo, A. Tiengo, S. Pedrazzoli, R. Seraglia, and M. Plebani, "Serum protein profiles of patients with pancreatic cancer and chronic pancreatitis: searching for a diagnostic protein pattern," *Rapid Communications in Mass Spectrometry*, vol. 15, pp. 2420-2425, 2001.

[83] J. Koopmann, Z. Zhang, N. White, J. Rosenzweig, N. Fedarko, S. Jagannath, M. I. Canto, C. J. Yeo, D. W. Chan, and M. Goggins, "Serum Diagnosis of Pancreatic Adenocarcinoma Using Surface-Enhanced Laser Desorption and Ionization Mass Spectrometry," *Clin Cancer Res*, vol. 10, pp. 860-868, 2004.

[84] J. M. Koomen, L. N. Shih, K. R. Coombes, D. Li, L.-c. Xiao, I. J. Fidler, J. L. Abbruzzese, and R. Kobayashi, "Plasma Protein Profiling for Diagnosis of Pancreatic Cancer Reveals the Presence of Host Response Proteins," *Clin Cancer Res*, vol. 11, pp. 1110-1118, 2005.

[85] J. M. Koomen, H. Zhao, D. Li, J. Abbruzzese, K. Baggerly, and R. Kobayashi, "Diagnostic protein discovery using proteolytic peptide targeting and identification," *Rapid Communications in Mass Spectrometry*, vol. 18, pp. 2537-48, 2004.

[86] Y. Won, H. Song, T. W. Kang, J. Kim, B. Han, and S. Lee, "Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons," *Proteomics*, vol. 3, pp. 2310 - 2316, 2003.

[87] R. Seraglia, E. Ragazzi, S. Vogliardi, G. Allegri, S. Pucciarelli, M. Agostini, M. Lise, D. Nitti, E. D. Urso, and P. Traldi, "Search of plasma markers for colorectal cancer by matrix-assisted laser desorption/ionization mass spectrometry," *Journal of Mass Spectrometry*, vol. 40, pp. 123-126, 2005.

[88] H. Bensmail, J. Golek, M. M. Moody, J. O. Semmes, and A. Haoudi, "A novel approach for clustering proteomics data using Bayesian fast Fourier transform," *Bioinformatics*, vol. 21, pp. 2210-2224, 2005.

[89] G. Ball, S. Mian, F. Holding, R. O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I. O. Ellis, C. Creaser, and R. C. Rees, "An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers," *Bioinformatics*, vol. 18, pp. 395-404, 2002.

[90] E. P. Diamandis, "Proteomic Patterns in Biological Fluids: Do They Represent the Future of Cancer Diagnostics?," *Clin Chem*, vol. 49, pp. 1272-1275, 2003.

[91] E. P. Diamandis, "Analysis of Serum Proteomic Patterns for Early Cancer Diagnosis: Drawing Attention to Potential Problems," *J Natl Cancer Inst*, vol. 96, pp. 353-356, 2004.

[92] E. P. Diamandis, "Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations," *Molecular & Cellular Proteomics*, vol. 3, pp. 367-78, 2004.

[93] K. A. Baggerly, J. S. Morris, and K. R. Coombes, "Reproducibility of SELDI-TOF protein patterns in serum: comparing data sets from different experiments," *Bioinformatics*, vol. 20, pp. 777-785, 2004.

[94] E. P. Diamandis and D.-E. van der Merwe, "Plasma Protein Profiling by Mass Spectrometry for Cancer Diagnosis: Opportunities and Limitations," *Clin Cancer Res*, vol. 11, pp. 963-965, 2005.

[95] E. F. I. Petricoin, D. K. Ornstein, C. P. Paweletz, A. Ardekani, P. S. Hackett, B. A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C. B. Simone, P. J. Levine, W. M. Linehan, M. R. Emmert-Buck, S. M. Steinberg, E. C. Kohn, and L. L. A., "Serum Proteomic Patterns for Detection of Prostate Cancer," *Journal of the National Cancer Institute*, vol. 94, pp. 1576-1578, 2002.

[96] K. A. Baggerly, J. S. Morris, S. R. Edmonson, and K. R. Coombes, "Signal in Noise: Evaluating Reported Reproducibility of Serum Proteomic Tests for Ovarian Cancer," *J Natl Cancer Inst*, vol. 97, pp. 307-309, 2005.

[97] A. I. Mehta, S. Ross, M. S. Lowenthal, V. Fusaro, D. A. Fishman, E. F. Petricoin, 3rd, and L. A. Liotta, "Biomarker amplification by serum carrier protein binding," *Disease Markers*, vol. 19, pp. 1-10, 2003.

[98] L. A. Liotta, M. Lowenthal, A. Mehta, T. P. Conrads, T. D. Veenstra, D. A. Fishman, and E. F. Petricoin, III, "Importance of Communication Between Producers and Consumers of Publicly Available Experimental Data," *J Natl Cancer Inst*, vol. 97, pp. 310-314, 2005.

[99] W. E. Grizzle, S. Meleth, E. F. Petricoin, and L. A. Liotta, "Clarification in the Point/Counterpoint Discussion Related to Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometric Identification of Patients with Adenocarcinomas of the Prostate * Proteomic Pattern Complexity Reveals a Rich and Uncharted Continent of Biomarkers," *Clin Chem*, vol. 50, pp. 1475-1477, 2004.

[100] O. J. Semmes, Z. Feng, B.-L. Adam, L. L. Banez, W. L. Bigbee, D. Campos, L. H. Cazares, D. W. Chan, W. E. Grizzle, E. Izbicka, J. Kagan, G. Malik, D. McLerran, J. W. Moul, A.

Partin, P. Prasanna, J. Rosenzweig, L. J. Sokoll, S. Srivastava, S. Srivastava, I. Thompson, M. J. Welsh, N. White, M. Winget, Y. Yasui, Z. Zhang, and L. Zhu, "Evaluation of Serum Protein Profiling by Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometry for the Detection of Prostate Cancer: I. Assessment of Platform Reproducibility," *Clin Chem*, vol. 51, pp. 102-112, 2005.

[101] W. E. Grizzle, B. L. Adam, W. L. Bigbee, T. P. Conrads, C. Carroll, Z. Feng, E. Izbicka, M. Jendoubi, D. Johnsey, J. Kagan, R. J. Leach, D. B. McCarthy, O. J. Semmes, S. Srivastava, S. Srivastava, I. M. Thompson, M. D. Thornquist, M. Verma, Z. Zhang, and Z. Zou, "Serum protein expression profiling for cancer detection: validation of a SELDI-based approach for prostate cancer," *Disease Markers*, vol. 19, pp. 185-195, 2003-2004.

[102] H. Shin and M. K. Markey, "A Machine Learning Perspective on the Development of Clinical Decision Support Systems Utilizing Mass Spectra of Blood Samples," *Journal of Biomedical Informatics*, vol. 39, pp. 227-248, 2006.

[103] K. R. Coombes, H. A. Fritsche, Jr., C. Clarke, J. N. Chen, K. A. Baggerly, J. S. Morris, L. C. Xiao, M. C. Hung, and H. M. Kuerer, "Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization," *Clinical Chemistry*, vol. 49, pp. 1615-23, 2003.

[104] M. S. Boguski and M. W. McIntosh, "Biomedical informatics for proteomics," *Nature*, vol. 422, pp. 233, 2003.

[105] Z. Feng, R. Prentice, and S. Srivastava, "Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective," *Pharmacogenomics*, vol. 5, pp. 709-19, 2004.

[106] J. Hu, K. R. Coombes, J. S. Morris, and K. A. Baggerly, "The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales," *Briefings in Functional Genomics and Proteomics*, vol. 3, pp. 322-331, 2005.

[107] S. de Waele and P. M. T. Broersen, "The Burg algorithm for segments," *IEEE Transactions on Signal Processing*, vol. 48, pp. 2876-2880, 2000.

[108] H. Shin, M. P. Sampat, J. Koomen, and M. K. Markey, "Statistical Characterization of Chemical Noise in MALDI TOF MS by Wavelet Analysis of Multiple Noise realizations," presented at American Medical Informatics Association (AMIA) Annual Symposium, Washington D.C., 2006.

[109] H. Shin, B. Sheu, and M. K. Markey, "Guilt-by-association Feature Selection Applied to Simulated Proteomic Data," presented at American Medical Informatics Association (AMIA) Annual Symposium, Washington D.C., 2005.

[110] H. Shin, B. Sheu, M. Joseph, and M. K. Markey, "Guilt-By-Association Feature Selection: Identifying Biomarkers from Proteomic Profiles," *Journal of Biomedical Informatics*, vol. submitted, submitted.

[111] L. Pusztai, B. W. Gregory, K. A. Baggerly, B. Peng, J. Koomen, H. M. Kuerer, F. J. Esteva, W. F. Symmans, P. Wagner, G. N. Hortobagyi, C. Laronga, O. J. Semmes, G. L. Wright, Jr., R. R. Drake, and A. Vlahou, "Pharmacoproteomic analysis of prechemotherapy and postchemotherapy plasma samples from patients receiving neoadjuvant or adjuvant chemotherapy for breast carcinoma," *Cancer*, vol. 100, pp. 1814-22, 2004.

[112] R. Aebersold and D. R. Goodlett, "Mass Spectrometry in Proteomics," *Chemical Review*, vol. 101, pp. 269-295, 2001.

[113] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, pp. 198-207, 2003.

[114] G. Siuzdak, *Mass spectrometry for biotechnology / Gary Siuzdak*. San Diego, Calif.: Academic Press, 1996.

[115] S. P. Gygi and R. Aebersold, "Mass spectrometry and proteomics," *Current Opinion in Chemical Biology*, vol. 4, pp. 489-494, 2000.

[116] J. R. Yates, 3rd, "Mass spectrometry. From genomics to proteomics," *Trends in Genetics.*, vol. 16, pp. 5-8, 2000.

[117] M. Mann, R. C. Hendrickson, and A. Pandey, "Analysis of proteins and protemes by mass spectrometry," *Annual Review of Biochemistry*, vol. 70, pp. 437-473, 2001.

[118] A. N. Krutchinsky and B. T. Chait, "On the nature of the chemical noise in MALDI mass spectra," *Journal of the American Society for Mass Spectrometry*, vol. 13, pp. 129-134, 2002.

[119] B. O. Keller and L. Li, "Discerning matrix-cluster peaks in matrix-assisted laser desorption/ionization time-of-flight mass spectra of dilute peptide mixtures," *Journal of the American Society for Mass Spectrometry*, vol. 11, pp. 88-93, 2000.

[120] T. W. Hutchens and T. T. Yip, "New Desorption Strategies for the Mass Spectrometric Analysis of Macromolecules," *Rapid Communications in Mass Spectrometry*, vol. 7, pp. 576-80, 1993.

[121] N. Tang, P. Tornatore, and S. R. Weinberger, "Current developments in seldi affinity tehcnology," *Mass spectrometry reviews*, vol. 1, pp. 34-44, 2004.

[122] M. Merchant and S. R. Weinberger, "Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry," *Electrophoresis*, vol. 21, pp. 1164-1177, 1999.

[123] W. Wu, W. Hu, and J. J. Kavanagh, "Proteomics in cancer research," *International Journal of Gynecological Cancer*, vol. 12, pp. 409-23, 2002.

[124] H. J. Issaq, T. D. Veenstra, T. P. Conrads, and D. Felschow, "The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification," *Biochemical & Biophysical Research Communications*, vol. 292, pp. 587-92, 2002.

[125] A. Vander, J. Sherman, and D. Luciano, *Human Physiology*, 8th ed. New York: McGraw-Hill, 2001.

[126] N. L. Anderson and N. G. Anderson, "The human plasma proteome: history, character, and diagnostic prospects.[erratum appears in Mol Cell Proteomics. 2003 Jan;2(1):50]," *Molecular & Cellular Proteomics*, vol. 1, pp. 845-67, 2002.

[127] K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M. C. Hung, and H. M. Kuerer, "Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform," M. D. Anderson Cancer Center, Houston 2004.

[128] G. A. Satten, S. Datta, H. Moura, A. R. Woolfitt, M. d. G. Carvalho, G. M. Carlone, B. K. De, A. Pavlopoulos, and J. R. Barr, "Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens," *Bioinformatics*, vol. 20, pp. 3128-3136, 2004.

[129] D. I. Malyarenko, W. E. Cooke, B.-L. Adam, G. Malik, H. Chen, E. R. Tracy, M. W. Trosset, M. Sasinowski, O. J. Semmes, and D. M. Manos, "Enhancement of Sensitivity and Resolution of Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometric Records for Serum Peptides Using Time-Series Analysis Techniques," *Clin Chem*, vol. 51, pp. 65-74, 2005.

[130] H. Shin, J. Koomen, K. A. Baggerly, and M. K. Markey, "Towards a Noise Model of MALDI TOF Spectra," presented at American Association for Cancer Research (AACR) Advances in Proteomics in Cancer Research, Key Biscayne, FL, 2004.

[131] F. R. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*. . New York: Springer-Verlag, 1985.

[132] M. Z. Wang, B. Howard, M. J. Campa, E. F. J. Patz, and F. M. C., "Analysis of human serum proteins by liquid phase isoelectric focusing and matrix-assisted laser desorption/ionization-mass spectrometry," *Proteomics*, vol. 3, pp. 1661-1666, 2003.

[133] M. Anderle, S. Roy, H. Lin, C. Becker, and K. Joho, "Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum," *Bioinformatics*, vol. 20, pp. 3575-82, 2004.

[134] X.-G. Shao, A. K.-M. Leung, and F.-T. Chau, "Wavelet: A New Trend in Chemistry," *Accounts of Chemical Research*, vol. 36, pp. 276 -283, 2003.

[135] V. J. Barclay and R. F. Bonner, "Application of Wavelet transforms to experimental spectra: smoothing, denoising, and data set compression," *Analytical Chemistry*, pp. 78-90, 1997.

[136] E. A. Robinson, *Statistical Communication and detection*. London: Griffin, 1967.

[137] J. Prados, A. Kalousis, J. C. Sanchez, L. Allard, O. Carrette, and M. Hilario, "Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents," *Proteomics*, vol. 4, pp. 2320-2332, 2004.

[138] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2000.

[139] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[140] E. T. Fung and C. Enderwick, "ProteinChip clinical proteomics: computational challenges and solutions," *Biotechniques. Suppl*, 2002.

[141] A. K. Jain, R. P. W. Duin, and M. Jianchang, "Statistical pattern recognition: a review," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, pp. 4, 2000.

[142] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, pp. 1437-1447, 2003.

[143] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, pp. 245-271, 1997.

[144] M. 4192Dash and H. Liu, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, 1997.

[145] A. Vlahou, P. F. Schellhammer, S. Mendrinos, K. Patel, F. I. Kondylis, L. Gong, S. Nasim, and G. L. Wright, Jr., "Development of a Novel Proteomic Approach for the Detection of Transitional Cell Carcinoma of the Bladder in Urine," *American Journal of Pathology*, vol. 158, pp. 1491-1502, 2001.

[146] T. M. Mitchell, *Machine Learning*. Boston: WCB/McGraw-Hill, 1997.

[147] C. E. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology.*, vol. 21, pp. 720-33, 1986.

[148] L. J. Lancashire, S. Mian, I. O. Ellis, R. C. Rees, and G. R. Ball, "Current Developments in the Analysis of Proteomic Data: Artificial Neural Network Data Mining Techniques for the Identification of Proteomic Biomarkers Related to Breast Cancer," *Current Proteomics*, vol. 2, pp. 15-29, 2005.

[149] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, 1 ed: Cambridge University Press, 2000.

[150] N. Cristianini and B. Scholkopf, "Support vector machines and kernel methods: the new generation of learning machines," in *AI Magazine*, 2002.

[151] M. Pontil and A. Verri, "Properties of Support Vector Machines," *Neural Comp.*, vol. 10, pp. 955-974, 1998.

[152] E. Bauer and R. Kohavi, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants," *Machine Learning*, vol. 36, pp. 105-139, 1999.

[153] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[154] N. Japkowicz, " Learning from Imbalanced Data Sets: a Comparison of Various Strategies.," AAAI Press, Menlo Park, CA, Technical Report WS-00-05, 2000.

[155] M. A. Maloof, "Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown.," Department of Computer Science, Georgetown University, Washington DC 2003.

[156] S. B. Kotsiantis and P. E. Pintelas, "Mixture of Expert Agents for Handling Imbalanced Data Sets," *Annals of Mathematics, Computing & Teleinformatics*, vol. 1, pp. 46-55, 2003.

[157] K. Orr, "Data quality and systems theory.," in *CACM*, vol. 41, 1998, pp. 66-71.

[158] T. C. Redman, "The Impact of Poor Data Quality on the Typical Enterprise.," in *CACM*, vol. 41, 1998, pp. 79-82.

[159] J. I. Maletic and A. Marcus, "Data cleansing: Beyond integrity analysis.," presented at Information Quality(IQ2000), Boston, MA, 2000.

[160] L. E. Dodd, R. F. Wagner, S. G. Armato, 3rd, M. F. McNitt-Gray, S. Beiden, H. P. Chan, D. Gur, G. McLennan, C. E. Metz, N. Petrick, B. Sahiner, J. Sayre, and G. Lung Image Database Consortium Research, "Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: contemporary research

topics relevant to the lung image database consortium," *Academic Radiology*, vol. 11, pp. 462-75, 2004.

[161] J. Han and M. Kamber, *Data Mining: Concepts and techniques*, 1st Edition ed. San Diego: Academic Press, 2001.

[162] C. E. Brodley and M. A. Friedl, "Identifying and eliminating mislabeled training instances.," presented at the 13th National Conference on Artificial Intelligence, Portland, OR, 1996.

[163] D. Gamberger, N. Lavrac, and C. Groselj, " Experiments with noise filtering in a medical domain.," presented at International Conference of Machine Learning (ICML'99), San Francisco, CA., 1999.

[164] B. Efron and G. Gong, "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *American Statistician*, vol. 37, pp. 36-48, 1983.

[165] B. Efron and R. Tibshirani, "Statistical Data Analysis in the Computer Age," *Science*, vol. 253, 1991.

[166] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall, 1993.

[167] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford University Press, 1995.

[168] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* Springer, 2002.

[169] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Boston: Academic Press, 1990.

[170] K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M.-C. Hung, and H. M. Kuerer, "Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform," *Proteomics*, vol. 5, pp. 4107-4117, 2005.

[171] V. P. Andreev, T. Rejtar, H. S. Chen, E. V. Moskovets, A. R. Ivanov, and B. L. Karger, "A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain," *Analytical chemistry*, vol. 75, pp. 6314-26, 2003.

[172] M. Statheropoulos, A. Pappa, P. Karamertzanis, and H. L. C. Meuzelaar, "Noise reduction of fast, repetitive GC/MS measurements using principal component analysis (PCA)," *Analytica Chimica Acta*, vol. 401, pp. 35-43, 1999.

[173] C. A. Hastings, S. M. Norton, and S. Roy, "New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data," *Rapid Communications in Mass Spectrometry*, vol. 16, pp. 462-467, 2002.

[174] V. Baranov, "Method for reducing chemical background in mass spectra," A. Biosystems, Ed., H01J049/42 ed. U.S.A.: MDS Inc., 2001.

[175] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*, Third ed: Prentice-Hall, 2000.

[176] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, pp. 716-723, 1974.

[177] S. Kullback, *Information theory and statistics*. New York: Wiely, 1959.

[178] K. P. Burnham and D. R. Anderson, "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods Research %R 10.1177/0049124104268644*, vol. 33, pp. 261-304, 2004.

[179] S. De Waele, "Automatic Inference from Finite Time Observations of Stationary Stochastic Signals," vol. Ph.D. Wageningen: Delft University of Technology, 2003, pp. 205.

[180] P. M. T. Broersen and S. de Waele, "Generating data with prescribed power spectral density," *Instrumentation and Measurement, IEEE Transactions on*, vol. 52, pp. 1061-1067, 2003.

[181] K. R. Coombes, J. M. koomenk, K. A. Baggerly, J. S. Morris, and R. Kobayashi, "Understanding the characteristics of mass spectrometry data through the use of simulation," The University of Texas M.D. Anderson Cancer Center, Houston 2004.

[182] R. C. Beavis and C. B.T., "Velocity distributions of intact high mass polypeptide molecule ions produced by matrix assisted laser desorption," vol. 181, pp. 479-484, 1991.

[183] P. Juhasz, M. L. Vestal, and S. A. Martin, "On the initial velocity of ions generated by matrix-assisted laser desorption ionization and its effect on the calibration of delayed extraction time-of-flight mass spectra," *Journal of the American Society for Mass Spectrometry*, vol. 8, pp. 209-217, 1997.

[184] D. W. Koppenaal, C. J. Barinaga, M. B. Denton, R. P. Sperline, G. M. Hieftje, G. D. Schilling, and F. J. Andrade, "MS Detector," in *Analytical Chemistry*, vol. 77, 2005, pp. 419A-427A.

[185] J. Ladislas Wiza, "Microchannel plate detectors," *Nuclear Instruments and Methods*, vol. 162, pp. 587-601, 1979.

[186] H. W. Ott, *Noise reduction techniques in electronic systems*, 2nd ed. New York: Wiely 1988.

[187] D. L. Donoho and I. M. Johnston, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, pp. 1200-1224, 1995.

[188] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrica*, vol. 81, pp. 425-455, 1994.

[189] S. G. Nikolov, H. Hutter, and M. Grasserbauer, "De-noising of SIMS images via wavelet shrinkage," *Chemometrics and Intelligent Laboratory Systems*, vol. 34, pp. 263-273, 1996.

[190] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, pp. 674-693, 1989.

[191] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising," *Wavelets and Statistics*, pp. 125-150, 1995.

[192] P. Vautrot, A. Ricordeau, and N. Bonnet, "Wavelet-based enoising from multiple noisy realizations: preliminary experiments," presented at SPIE: the International Society for Optical Engineering, San Jose, California, 2000.

[193] J. C. Goswami and A. K. Chan, "Fundamentals of Wavelets: Theory," *Applications, and Algorithms (John Wiley & Sons, NY, USA)*, 1999.

[194] C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to wavelets and wavelet transforms: a primer*. Upper Saddle River, New Jersey: Prentice-Hall Inc., 1998.

[195] S. Mallat, "A Wavelet Tour of Signal Processing (ed.)," *New York: Academic*, 1999.

[196] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," *Wavelets and Statistics*, vol. 103, pp. 281–299, 1995.

[197] J. C. Pesquet, H. Krim, and H. Carfantan, "Time-invariant orthonormal wavelet representations," *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, vol. 44, pp. 1964-1970, 1996.

[198] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells, "Noise reduction using an undecimated discrete wavelet transform," *IEEE Signal Processing Letters*, vol. 3, pp. 10-12, 1996.

[199] D. L. Donoho, "De-noising by soft-thresholding," *Information Theory, IEEE Transactions on*, vol. 41, pp. 613-627, 1995.

[200] L. Debnath, *Wavelets and Signal Processing*. Boston: Birkhauser, 2003.

[201] W. Dumouchel and F. O'Brien, "Integrating a robust option into a multiple regression computing environment," *Ima Volumes In Mathematics And Its Applications*, pp. 41-48, 1992.

[202] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton Univesity Press, 1961.

[203] D. Koller and M. Sahami, "Towards Optimal Feature Selection," presented at Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 1996.

[204] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*: Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[205] M. Mitchell, *An Introduction to Genetic Algorithms*: Bradford Book, 1998.

[206] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," in *Department of Computer Science*, vol. Doctor of Philosophy. Hamilton, New Zealand: The University of Waikato, 1999, pp. 178.

[207] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 301-312, 2002.

[208] B. King, "Step-Wise Clustering Procedures," *Journal of the American Statistical Association*, vol. 62, pp. 86-101, 1967.

[209] X. Rui and D. Wunsch, II, "Survey of clustering algorithms," *Neural Networks, IEEE Transactions on*, vol. 16, pp. 645-678, 2005.

[210] S. Sharma, *Applied Multivariate Techniques*. New York: John Wiley & Sons, 1995.

[211] F. J. Schweigert, K. Wirth, and J. Raila, "Characterization of the microheterogeneity of transthyretin in plasma and urine using SELDI-TOF-MS immunoassay," *Proteome Science*, vol. 2, 2004.

[212] E. F. Petricoin Iii, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, pp. 572-577, 2002.

[213] "ExPASy molecular biology server. Available at http://us.expasy.org."

[214] "UCSC proteom browser. Available at http://genome.ucsc.edu/google/goldenPath/help/pbTracksHelpFiles/pbpepPi.shtml."

[215] S. Struyf, P. Menten, J. P. Lenaerts, W. Put, A. D'Haese, E. De Clercq, D. Schols, P. Proost, and J. Van Damme, "Diverging binding capacities of natural LD78 isoforms of macrophage inflammatory protein-1 to the CC chemokine receptors 1, 3 and 5 affect their anti-HIV-1 activity and chemotactic potencies for neutrophils and eosinophils," *Eur. J. Immunol*, vol. 31, pp. 2170-2178, 2001.

[216] R. Hoffmann, T. Seidl, and M. Dugas, "Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis," *Genome Biology*, vol. 3, 2002.

[217] L.-X. Qin, K. F. Kerr, and C. M. o. t. T. R. Consortium, "Empirical evaluation of data transformations and ranking statistics for microarray analysis," *Nucleic Acids Research*, vol. 32, pp. 5471-5479, 2004.

Vita

Hyunjin Shin was born in Seoul, South Korea on July 6th, 1972 as the son of Young-Kyun Shin and Bookang Rhee. He entered Seoul National University, Seoul, South Korea in 1992. He completed his B.S. degree in the School of Electrical Engineering at Seoul National University in August 1998. He began his M.S. degree in the Department of Electrical and Computer Engineering at the University of Texas at Austin in September 1999 and received his M.S. degree in December 2001 under Dr. Jonathan W. Valvano's supervision. The master thesis's topic was the development of a real-time data acquisition system on Windows operating systems. He joined the Biomedical Informatics Lab at The University of Texas at Austin in October 2002 and has developed algorithms for biomarker identification using MALDI TOF mass spectrometry under Dr. Mia K. Markey and Dr. Jonathan W. Valvano's supervision.

Permanent address: Royal Count East #102, Gumi-dong 276, Pundang-gu, Sungnam-si, Kyunggi-do, South Korea 463-802

This dissertation was typed by the author.